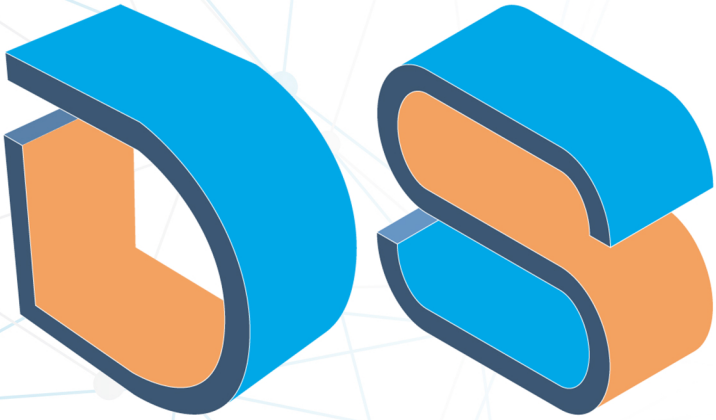


POLIMI
DATA SCIENTISTS

MIDA 1
Course Notes

Edited by:
Marco Varrone

These notes have been made thanks to the effort of Polimi Data Scientists staff.

Are you interested in Data Science activities?

Follow PoliMi Data Scientists on Facebook!

Polimi Data Scientist is a community of students and Alumni of Politecnico di Milano.

We organize events and activities related to Artificial Intelligence and Machine Learning, our aim is to create a strong and passionate community about Data Science at Politecnico di Milano.

Do you want to learn more?
Visit our website and join our Telegram Group! !

Credits

The following notes have been written by the Polimi Data Scientists student association by combining Prof. Bittanti's lectures and notes with content from the *Identificazione dei Modelli e Analisi dei Dati 1 2010-2011* notes by Stefano Invernizzi.

They are meant as a support for the students following the course and they should not be considered as a replacement for the professor's lectures or the book suggested in the course bibliography.

Contents

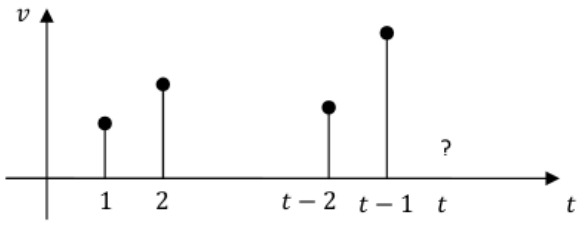
Part I

Prediction

Chapter 1

The prediction problem

Let's consider a sequence ordered by t: $v(t-1), v(t-2), ...$ Can we predict the value of $v(t)$?



Suppose that we don't know how the data has been generated.

1.1 Symbolism

The unknown datum $v(t)$ represents the value that v will assume at time t. However, we need to distinguish between the actual value in that moment and the one resulting from out estimation, which is $\hat{v}(t)$.
The notation $\hat{v}(t|t-1)$ describes the estimate of v at time t, given the past values at time $t-1$, $t-2$, ... that are all the past values up to $t-1$. In this case, it is called 1-step-ahead predictor.

1.2 The linear predictor

We can build a linear predictor, which is a predictor obtained by computing $\hat{v}(t|t-1)$ as a linear combination of the samples. We can build a finite memory linear predictor: a predictor computed as a linear combination of the last n samples:

$$\hat{v}(t|t-1) = a_1 \cdot v(t-1) + a_2 \cdot v(t-2) + ... + a_n \cdot v(t-n)$$

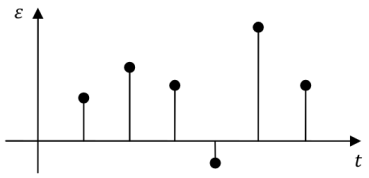
We then need to find a criterion to estimate the values of $a_1, a_2, ..., a_n$.
After fixing all the parameters, we can compute the prediction error

$$\varepsilon = v(t) - \hat{v}(t|t-1)$$

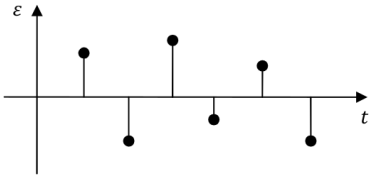
Because $v(\cdot)$ and $\hat{v}(\cdot)$ change over time, $\varepsilon(\cdot)$ changes too. More specifically, since $v(\cdot)$ is a stochastic process, $\varepsilon(\cdot)$ is stochastic, too.
Let's analyze the properties of $\varepsilon(\cdot)$ leading to a good predictor. The best case would obviously be an error always equal to zero. This is very often impossible. But we can also consider the following two cases:

- The prediction error as mean different from zero. We can easily obtain a better predictor by shifting the values of $v(\cdot)$ by the mean to obtain a prediction error with zero mean.

- The prediction error has zero mean, but we can see that its sign changes at each time step. We can obtain a better predictor because at each following time step we know that the prediction is wrong because of either an overestimation or an underestimation.



(a) Sequence with positive mean



(b) Sequence with alternating sign

Our aim is then to obtain a fully unpredictable error. This type of error is called white noise

$$\varepsilon(\cdot) \sim WN(0, \lambda^2)$$

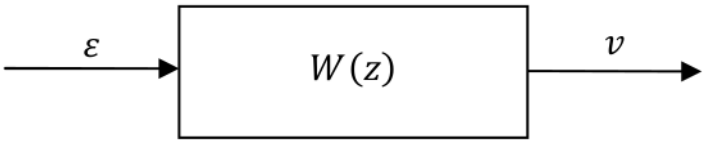
Where the two parameters are, respectively, the mean and the variance of $\varepsilon(\cdot)$.
Let's suppose that $\varepsilon(\cdot)$ is indeed a white noise. From the definition of prediction error

$$\varepsilon(\cdot) = v(t) - \hat{v}(t|t-1)$$

We can then obtain

$$v(t) = \hat{v}(t|t-1) + \varepsilon(t) = a_1 \cdot v(t-1) + ... + a_n \cdot v(t-n) + \varepsilon(t)$$

The equation above is a discrete difference equation having $v(\cdot)$ as a unknown variable, which can be seen as the output of a linear system with input $\varepsilon(\cdot)$.



Starting from the equation we can obtain the transfer function *W(z)* by introducing the operator *z*. Note that

$$\begin{aligned}
z \cdot v(t) &= v(t+1) && \text{(1-step-ahead forward operator)} \\
z^{-1} \cdot v(t) &= v(t-1) && \text{(1-step-ahead backward operator)} \\
z^{n} \cdot v(t) &= v(t+n) && (n\text{-step-ahead forward operator)} \\
z^{-n} \cdot v(t) &= v(t-n) && (n\text{-step-ahead backward operator)}
\end{aligned}$$

We can hence have:

$$\begin{aligned}
v(t) &= a_1 z^{-1} v(t) + a_2 z^{-2} v(t) + ... + a_n z^{-n} v(t) + \varepsilon(t) \quad \Rightarrow \\
\Rightarrow \quad & (1 - a_1 z^{-1} - a_2 z^{-2} - ... - a_n z^{-n}) v(t) = \varepsilon(t) \quad \Rightarrow \\
\Rightarrow \quad W(z) &= \frac{v(t)}{\varepsilon(t)} = \frac{1}{1 - a_1 z^{-1} - a_2 z^{-2} - ... - a_n z^{-n}} = \frac{z^n}{z^n - a_1 z^{n-1} - ... - a_n}
\end{aligned}$$

Note that z is a complex variable.
We can now compute the zeros and poles of the transfer function:

- the zeros are the values for which the numerator of *W(z)* is equal to zero. The result is n zeros at the origin.
- the poles are obtained by forcing the denominator of *W(z)* to zero. The result is n poles that can be in any position of the space, with the constraint that if there is a complex pole, there is also its conjugate.

In conclusion, to obtain a good linear predictor, we need to describe the exact signal as the output of a system with a transfer function described as above and with a white noise as input.

Chapter 2

Random concepts

2.1 Random variable

A random variable $v(s)$ is a real function of a random event, associated to the outcome s of a random experiment.

Mean, average or expected value

The mean of a random variable is always a real number

$$E[v] = m = \bar{v}$$

It has the following property: $E[a_1 \cdot v_1 + a_2 \cdot v_2] = a_1 E[v_1] + a_2 E[v_2]$, with v_1, v_2 random variables and a_1, a_2 real numbers.

Variance

The variance of a random variable is a non-negative real number

$$Var[v] = E[(v - E[v])^2] = \lambda^2 = \sigma^2 = \mu^2$$

For two variables v_1 and v_2, the square root of the variance (**standard deviation**) can be expressed respectively as λ_{11} and λ_{22}.

Covariance (or cross-variance)

Given to random variables v_1 and v_2, the covariance between these two variables is

$$\lambda_{12} = \lambda_{21} = E[(v_1 - E[v_1])(v_2 - E[v_2])]$$

Property

In case of a random variable with Gaussian distribution $v \sim G(m, \lambda^2)$, in 95% (99%) of cases, v will take value in the interval $m \pm 2\lambda$ ($m \pm 3\lambda$).

2.2 Random vectors

A random vector is a set of random variables. It is represented as a column vector

$$v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

Mean

The mean value of a random vector is the vector of the means of its values

$$E[v] = \begin{bmatrix} E[v_1] \\ E[v_2] \end{bmatrix}$$

Variance

The variance of a random vector of length n is a symmetric matrix of size $n \times n$. For $n = 2$

$$Var[v] = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix}$$

The i^{th} element of the main diagonal is the variance of the i^{th} component of the random vector, while the (i, j) positions contain the value of the cross-variance of the i^{th} and j^{th} component of the random vector.
Some remarks:

$$\Delta v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} - \begin{bmatrix} E[v_1] \\ E[v_2] \end{bmatrix}$$

Then the variance can be written as

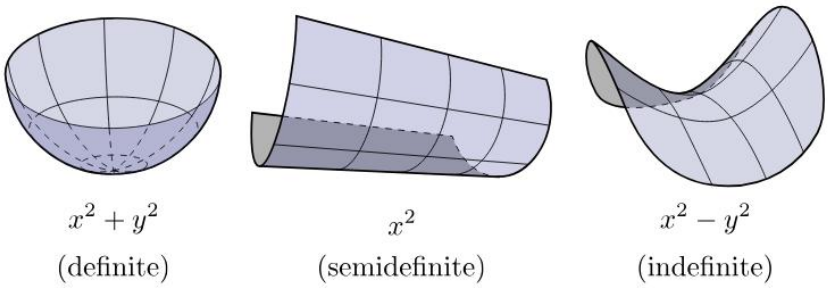
$$Var[v] = E[\Delta v \Delta v']$$

Properties

$Var[v]$ is a positive semi-definite matrix.
This means that the quadratic form is a positive semi-definite matrix A

$$f(x_1, x_2) = \begin{bmatrix} x_1 & x_2 \end{bmatrix} A \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

For $n = 2$, we note that A is a conic function of the following type.



$x^2 + y^2$ (definite) x^2 (semidefinite) $x^2 - y^2$ (indefinite)

$A \geq 0 \equiv$ positive semi-definite and, thus, $\det A \geq 0$.
Which implies that

$$\lambda_{11}\lambda_{22} - \lambda_{12}\lambda_{21} \geq 0 \quad \rightarrow \quad \lambda_{11}\lambda_{22} - \lambda_{12}^2 \geq 0$$

Covariance coefficient

The covariance coefficient of a random vector of length 2 is

$$\rho = \frac{\lambda_{12}}{\sqrt{\lambda_{11}}\sqrt{\lambda_{22}}}$$

Given that the variance is a positive semi-definite matrix we can conclude that

$$|\rho| \leq 1$$

In particular, if $\rho = 0$, then v_1 and v_2 are uncorrelated.

Example
Given the random variables v_1, v_2 related by $v_2 = \alpha v_1$.
Given that $E[v_1] = 0$ and $Var[v_1] = \lambda_{11}$.
Then

$$E[v_2] = E[\alpha v_1] = \alpha E[v_1] = 0$$

$$\lambda_{12} = E[v_1 v_2] = E[v_1 \alpha v_1] = \alpha E[v_1^2] = \alpha \lambda_{11}$$

$$\lambda_{22} = E[(\alpha v_1)^2] = \alpha^2 E[v_1^2] = \alpha^2 \lambda_{11}$$

$$\rho = \frac{\lambda_{12}}{\sqrt{\lambda_{11}}\sqrt{\lambda_{22}}} = \frac{\alpha\lambda_{11}}{\sqrt{\lambda_{11}}\sqrt{\alpha^2\lambda_{11}}} = \frac{\alpha}{|\alpha|} = \begin{cases} +1 & \text{if } \alpha > 0 \\ -1 & \text{if } \alpha < 0 \end{cases}$$

Example extended
Consider the following case:

$$v_2 = \alpha v_1 + e$$

Where e is a random variable with 0 mean and variance μ^2 (noise).
If we want to find ρ as a function of μ^2, we obtain that if $\mu^2 = 0$, then $|\rho| = 1$ and then $|\rho|$ starts to decrease until it reaches 0 as μ^2 increases.

2.3 Stochastic (or random) process

A random process is a countable set of random variables. Hence, a random variable is usually indexed by t and it is represented as $v(t)$.
However, this notation doesn't highlight the fact that it is a function depending also on the outcome s of a random experiment. For this reason, the following notation can be used:

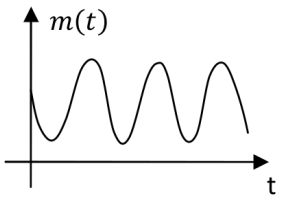
$$v(t, s)$$

By fixing the time $\bar{t}$ we obtain a random variable $v(\bar{t}, s) = v(\bar{t}, \cdot)$.
By fixing the outcome $\bar{s}$ we obtain the process realization $v(t, \bar{s}) = v(\cdot, s)$.

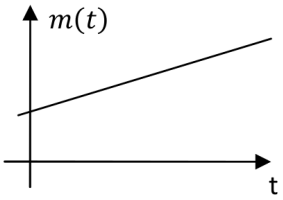
Mean

$$m(t) = E_s[v(t, s)] = E[v(t)]$$

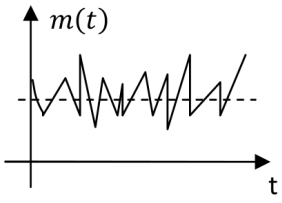
The mean value doesn't depend on s. Thus, the mean is computed over all possible outcomes. It can change over time:



(a) Periodic



(b) Linear



(c) Fluctuating

Variance

$$Var[v(t)] = E[(v(t) - m(t))^2] = \lambda^2(t)$$

Covariance (or cross-variance)

The covariance is obtained by considering two values of v at different times t_1 and t_2.

$$\gamma(t_1, t_2) = E[(v(t_1) - m(t_1)) \cdot (v(t_2) - m(t_2))]$$

2.4 Stationary process

A stationary process is a random process in which:

1. The mean value $m(t)$ is constant
2. The variance $\lambda^2(t)$ is constant
3. The covariance $\gamma(t_1, t_2)$ depends only on $\tau = t_2 - t_1$.

The $\gamma(t_1, t_2)$ notation is called double index, while the $\gamma(\tau)$ notation is called single index.

Properties

1. The covariance at $\tau = 0$ is equal to the variance of the process at time t

$$\gamma(0) = \gamma(t,t) = E[(v(t) - m(t))(v(t) - m(t))] = E[(v(t) - m(t))^2] = \lambda^2$$

   Hence $\gamma(0) \geq 0$

2. γ is an even function of τ

$$\gamma(t_1, t_2) = \gamma(t_2, t_1) \quad \Rightarrow \quad \gamma(-\tau) = \gamma(+\tau)$$

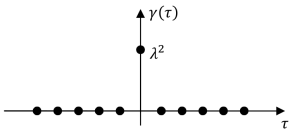
3. The absolute value of the covariance is never greater than $\gamma(0)$

$$|\gamma(\tau)| \leq \gamma(0)$$

2.4.1 White noise

The white noise $\eta(t) \sim WN(m, \lambda^2)$ can now be defined as a stationary process with:

1. $E[v(t)] = 0, \quad \forall t$
2. $Var[v(t)] = \lambda^2, \quad \forall t$
3. $\gamma(\tau) = \begin{cases} 0 & \text{if } \tau \neq 0 \\ \lambda^2 & \text{if } \tau = 0 \end{cases}$



The fact that the covariance function is zero everywhere except for the origin means that the notion of the past is not informative to know the future (**whiteness property**).

Chapter 3

AR, MA and ARMA Processes

3.1 MA Processes

3.1.1 MA(1) process

Given $\eta(n) \sim WN(0, \lambda^2)$ and a process with the following behavior:

$$v(t) = c_0\eta(t) + c_1\eta(t-1), \qquad c_0, c_1 \in \mathbb{R}$$

This type of process is called MA(1), that is Moving Average of order 1.

Mean value

$$E[v(t)] = E[c_0\eta(t)] + E[c_1\eta(t-1)] = c_0 \cdot 0 + c_1 \cdot 0 = 0$$

Variance

$$\begin{aligned} Var[v(t)] &= E[(v(t) - E[v(t)])^2] \\ &= E[v(t)^2] \\ &= E[(c_0\eta(t) + c_1\eta(t-1))^2] \\ &= E[c_0^2\eta(t)^2] + E[c_1^2\eta(t-1)^2] + E[2 \cdot c_0c_1\eta(t)\eta(t-1)] \\ &= c_0^2\lambda^2 + c_1^2\lambda^2 + 0 \\ &= (c_0^2 + c_1^2)\lambda^2 \end{aligned}$$

Covariance

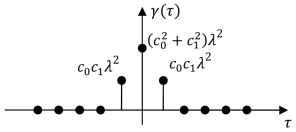
$$\begin{aligned} \gamma(t_1, t_2) &= E[(v(t_1) - E[v(t_1)])(v(t_2) - E[v(t_2)])] \\ &= E[(c_0\eta(t_1) + c_1\eta(t_1-1))(c_0\eta(t_2) + c_1\eta(t_2-1))] \\ &= c_0^2E[\eta(t_1)\eta(t_2)] + c_1^2E[\eta(t_1-1)\eta(t_2-1)] + c_1c_0E[\eta(t_1-1)\eta(t_2)] + c_0c_1E[\eta(t_1)\eta(t_2-1)] \end{aligned}$$

By considering that the noise at time t is completely uncorrelated with the noise at time *t - 1, t + 1, t - 2, t + 2,...* we can distinguish:

1. If $t_2 = t_1 \pm 1 : \gamma(t, t+1) = \gamma(t-1, t) = 0 + 0 + c_1c_0\lambda^2 + 0 = c_0c_1\lambda^2$
2. If $t_2 = t_1 \pm 2 : \gamma(t, t+2) = \gamma(t-2, t) = 0$
3. If $t_2 = t_1 \pm 3 : \gamma(t, t+3) = \gamma(t-3, t) = 0$
4. ...

In conclusion we obtain a stationary process, but in this case, it is not a white noise, because the covariance is not 0 for the values ± 1.
In summary:

- $E[v(t)] = 0, \quad \forall t$
- $Var[v(t)] = (c_0^2 + c_1^2)\lambda^2, \quad \forall t$
- $\gamma(t_1, t_2) = \begin{cases} 0 & \text{if } |t_1 - t_2 \\ c_0 c_1 \gamma^2 & \text{if } t_2 = t_1 \pm 1 \end{cases}$



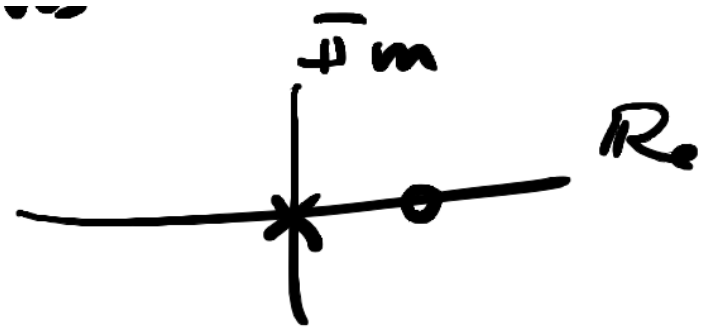
By moving to the z domain we obtain

$$v(t) = c_0\eta(t) + c_1\eta(t-1) = c_0\eta(t) + c_1 z^{-1}\eta(t) = (c_0 + c_1 z^{-1})\eta(t)$$

Then the transfer function $W(z)$ is

$$W(z) = c_0 + c_1 z^{-1} = c_0 + c_1 \frac{1}{z} = \frac{c_0 z + c_1}{z}$$

We have one pole at $z = 0$ and one zero at $c_0 z + c_1 = 0$



3.1.2 MA(n) process

We can extend the previous reasoning to a MA(n) process:

$$v(t) = c_0\eta(t) + c_1\eta(t-1) + ... + c_n\eta(t-n)$$

$$v(t) = (c_0 + c_1 z^{-1} + ... + c_n z^{-n})\eta(t)$$

The transfer function is, thus:

$$W(z) = c_0 + c_1 z^{-1} + ... + c_n z^{-n} = \frac{c_0 z^n + c_1 z^{n-1} + ... + c_n}{z^n}$$

Resulting in n zeros depending on the values of the c coefficients and n poles all in the origin. The class of MA processes is a class of stationary processes, with the following characteristics.

- $E[v(t)] = 0$
- $Var[v(t)] = (c_0^2 + c_1^2 + ... + c_n^2)\lambda^2$
- $\gamma(1) = (c_0 c_1 + c_1 c_2 + ... + c_{n-1} c_n)\lambda^2$
- $\gamma(2) = (c_0 c_2 + c_1 c_3 + ... + c_{n-2} c_n)\lambda^2$
- ...
- $\gamma(\pm n) = c_0 c_n \lambda^2$
- $\gamma(\pm k) = 0, \quad k > n$

3.1.3 MA(∞) process

Let's now consider the case in which the output signal is a linear combination over an infinite number of time steps.

$$v(t) = c_0\eta(t) + c_1\eta(t-1) + ... + c_{n-1}\eta(t-n+1) + c_n\eta(t-n) + c_{n+1}\eta(t-n-1) + ...$$

The variance is:

$$Var[v(t)] = (c_0^2 + c_1^2 + c_2^2 + ...)\lambda^2$$

The values of the covariance are the same as in the MA(n) process, but in this case there is no value of $k > n$ for which the covariance is 0.
For the variance to be finite, we need to verify that the following series is finite:

$$c_0^2 + c_1^2 + ... = \sum_{i=0}^{+\infty} c_i^2$$

If so, $\gamma(\tau)$ is guaranteed to be always finite because of the relation $|\gamma(\tau)| \leq \gamma(0)$.

3.2 AR Processes

3.2.1 AR(1) process

The behavior of an AR(1) process with $\eta \sim WN(0, \lambda^2)$ is described as follows:

$$v(t) = av(t-1) + \eta(t)$$

This process can be analyzed by using three different methods.

First method

We can observe that an AR(1) process can be expressed as a type of MA(∞).

$$\begin{aligned} v(t) &= av(t-1) + \eta(t) = a[av(t-2) + \eta(t-1)] + \eta(t) \\ &= a^2[av(t-3] + \eta(t-2)] + a\eta(t-1) + \eta(t) \\ &= ... \\ &= \underbrace{\eta(t) + a\eta(t-1) + a^2\eta(t-2) + ...}_{\text{MA}(\infty)} + a^n v(t-n) \end{aligned}$$

$v(t)$ is hence an MA(∞) with $c_i = a^i$ only if the last term is brought to zero. This is achieved by applying the constraints $|a| < 1$, for which $a^n v(t-n) \to 0$.
We also need to check if the variance is finite. As previously seen we need to compute:

$$\sum_{i=0}^{+\infty} c_i^2 = \sum_{i=0}^{+\infty} a^{2i}$$

which is a geometric series, convergent if $|a| < 1$. Under this hypothesis, the series is:

$$\sum_{i=0}^{+\infty} a^{2i} = \frac{1}{1-a^2}$$

Finally, if $|a| < 1$, then $v(t)$ is a stationary process with variance:

$$Var[v(t)] = \gamma(0) = [c_0^2 + c_1^2 + ... + c_n^2 + ...]\lambda^2 = \left(\sum_{i=0}^{+\infty} a^{2i}\right)\lambda^2 = \frac{\lambda^2}{1-a^2}$$

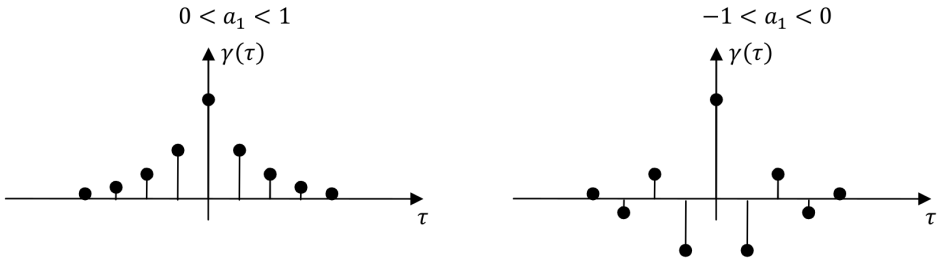
And covariance:

$$\gamma(1) = (c_0c_1 + c_1c_2 + c_2c_3 + ...)\lambda^2 = (a + aa^2 + a^2a^3 + ...)\lambda^2 = a(1 + a^2 + a^4 + ...)\lambda^2 = a\frac{1}{1-a^2}\lambda^2$$

$$\gamma(2) = (c_0c_2 + c_1c_3 + ...)\lambda^2 = a^2\frac{1}{1-a^2}\lambda^2$$

In general:

$$\gamma(\tau) = a^{|\tau|}\gamma(0) = a^{|\tau|}\frac{\lambda^2}{1-a^2}$$



Second method: Yule-Walker equations

Another possible method makes use of the Yule-Walker equations.

$$Var[v(t)] = E[v(t)^2] = E[(av(t-1) + \eta(t))^2] = a^2E[v(t-1)^2] + E[\eta(t)^2] + 2aE[v(t-1)\eta(t)]$$

Where $E[v(t-1)\eta(t)]$ is the correlation between $v(t-1)$ and $\eta(t)$. We know that $v(t-1)$ depends on $\eta(t-1), \eta(t-2), ...$ but given that $\eta(\cdot)$ is a white noise, it is not correlated with any other values at previous times. Thus, $v(t-1)$ and $\eta(t)$ are uncorrelated.
Furthermore, $v(\cdot)$ is a stationary process, so $Var[v(t)] = Var[t-1] = \gamma(0)$.
We obtain, from the first equation:

$$Var[v(t)] = \gamma(0) = a^2\gamma(0) + \lambda^2 + 0 \quad \rightarrow \quad \gamma(0) = \frac{\lambda^2}{1-a^2}$$

Let's now consider $\gamma(\tau)$. We start from the usual expression for $v(t)$ and we multiply each term by $v(t-\tau)$

$$v(t)v(t-\tau) = av(t-1)v(t-\tau) + \eta(t)v(t-\tau)$$

We compute the covariance.

$$\underbrace{E[v(t)v(t-\tau)]}_{\gamma(\tau)} = a\,\underbrace{E[v(t-1)v(t-\tau)]}_{\gamma(\tau-1)} + E[\eta(t)v(t-\tau)]$$

For the same reasoning above, $v(t-\tau)$ depends on $\eta(t-\tau), \eta(t-\tau-1), ...$, so if $\tau > 0$, $v(t-\tau)$ and $\eta(t)$ are uncorrelated.

$$\gamma(\tau) = a\gamma(\tau-1), \quad \forall\tau > 0$$

By combining the two results we obtain the Yule-Walker equations:

$$\begin{cases} \gamma(0)\frac{1}{1-a^2}\lambda^2 \\ \gamma(\tau) = a\gamma(\tau-1), \quad \forall\tau > 0 \end{cases}$$

And we obtain the same expression of the first method

$$\gamma(\tau) = \frac{a^{|\tau|}}{1-a^2}\lambda^2$$

Third method: long division

With this methods, we aim at expressing the transfer function in the form:

$$W(z) = c_0 + c_1 z^{-1} + c_2 z^{-2} + ...$$

From the behavior description of the system we can obtain the expression of the transfer function of the AR(1).

$$v(t) = av(t-1) + \eta(t) = az^{-1}v(t) + \eta(t)$$

$$W(z) = \frac{v(t)}{\eta(t)} = \frac{1}{1 - az^{-1}} = \frac{z}{z - a}$$

We can perform the long division between the numerator and the denominator of $W(z)$.

$$\begin{array}{llll} 1 & & & 1 - az^{-1} \\ \hline 1 & -az^{-1} & & 1 + az^{-1} \\ \hline / & az^{-1} & & \\ & az^{-1} & -a^2 z^{-2} & \\ \hline & / & a^2 z^{-2} & \end{array} \tag{3.1}$$

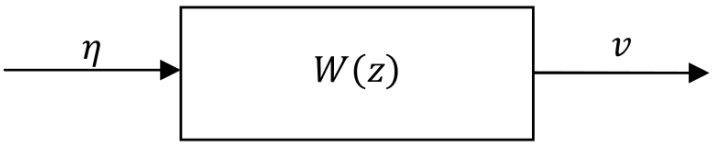
And so on, we obtain:

$$W(z) = 1 + az^{-1} + a^2 z^{-2} + ...$$

Which is equal to the transfer function of a MA(∞) process, by setting $c_i = a^i$:

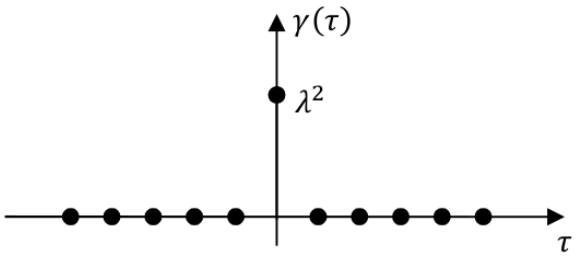
$$W(z) = c_0 + c_1 z^{-1} + c_2 z^{-2} + ...$$

In general, given any transfer function, we can always write it as above, with the coefficients found with the long division algorithm (that now we will call with the generic notation w_i).
Let's analyze the meaning of these coefficients. By representing the system with a block scheme



Assume the noise to be a discrete time impulse:

$$\eta(t) = imp(t) = \begin{cases} 1, & t = 0 \\ 0, & t \neq 0 \end{cases}$$



The corresponding output $v(t)$ is the impulse response:

$$v(t) = w_0\eta(t) + w_1\eta(t-1) + ...$$

For $t = 0$ we have:

$$v(0) = w_0\eta(0) + w_1\eta(-1) + ...$$

But if η is an impulse, we have that $v(0) = w_0$.
For $t = 1$ we have:

$$v(1) = w_0\eta(1) + w_1\eta(0) + ... = w_1$$

And so on. Thus, in the transfer function

$$W(z) = w_0 + w_1 z^{-1} + ...$$

w_0 can be interpreted as the impulse response of $W(z)$ at time $t = 0$, w_1 as the impulse response of $W(z)$ at time $t = 1$, ...
As always, $v(t)$ is stationary when $w_0^2 + w_1^2 + ...$ is finite, which can be ensured by checking that the system is stable (i.e. all the poles of $W(z)$ have absolute value < 1).

3.2.2 AR(n) process

Let's consider a generic autoregressive model of order n.

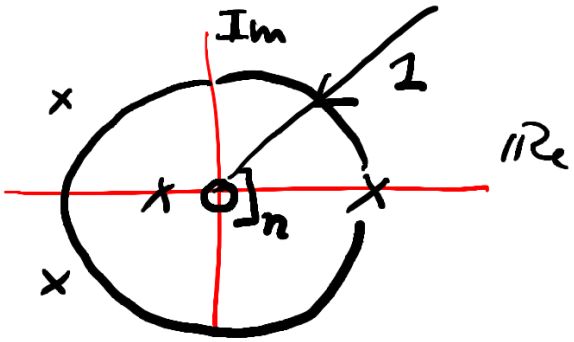
$$v(t) = a_1 v(t-1) + a_2 v(t-2) + ... + a_n v(t-n) + \eta(t)$$

We can obtain the value of the transfer function from the model in operator form, as we have done for the AR(1).

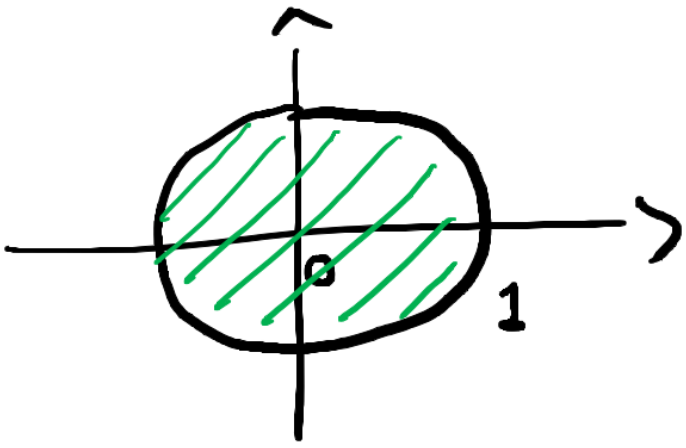
$$v(t) = a_1 z^{-1} v(t) + a_2 z^{-2} v(t) + ... + a_n z^{-n} v(t) + \eta(t)$$
$$(1 - a_1 z^{-1} - a_2 z^{-2} - ... - a_n z^{-n}) v(t) = A(z) v(t) = \eta(t)$$
$$v(t) = W(z)\eta(t) \quad \rightarrow \quad W(z) = \frac{1}{A(z)} = \frac{z^n}{z^n - a_1 z^{-1} - ... - a_n}$$

By now seeing z as a complex variable, $W(z)$ is the transfer function from $\eta(t)$ to $v(t)$, with:

- $z^n = 0$: n zeros all in the origin
- $z^n - a_i z^{n-1} - ... - a_n = 0$: n poles depending on the values of a_i.



Note that if all poles are located inside the unit disk, then $v(\cdot)$ is stationary.



3.3 ARMA processes

3.3.1 ARMA(n_a, n_c) process

It is are a family of processes that includes all the AR and MA processes:

$$v(t) = \underbrace{a_1 v(t-1) + a_2 v(t-2) + ... + a_{n_a} v(t-n_a)}_{\text{AR}} + \underbrace{c_0 \eta(t) + c_1 \eta(t) + ... + c_{n_c} \eta(t-n_c)}_{\text{MA}}$$

Where n_a is the order of the AR part and n_c is the order of the MA part.
The transfer function can be computed as follows:

$$v(t) = a_1 z^{-1} v(t) + ... + a_{n_a} z^{-n_a} v(t) + c_0 \eta(t) + c_1 z^{-1} \eta(t) + ... + c_{n_c} z^{-n_c} \eta(t)$$

$$(1 - a_1 z^{-1} - ... - a_{n_a} z^{-n_a}) v(t) = (c_0 + c_1 z^{-1} + ... + c_{n_c} z^{-n_c}) \eta(t) \quad \rightarrow \quad A(z) v(t) = C(z) v(t)$$

$$W(z) = \frac{C(z)}{A(z)} = \frac{c_0 + c_1 z^{-1} + ... + c_{n_c} z^{-n_c}}{1 - a_1 z^{-1} - ... - a_{n_a} z^{-n_a}}$$

It is important to distinguish between the concept of ARMA model, which is the system described by the first equation, from the concept of ARMA process, which is the the process $v(\cdot)$ generated from the ARMA model, only in the case the generated process is stationary.
Again, the ARMA process is stationary if all the poles of $W(z)$ are inside the unit disk.

Chapter 4

Frequency domain

4.1 Spectrum

Assume $v(\cdot)$ to be a stationary stochastic process, and $\gamma(\tau)$ to be its covariance function. The spectrum is defined as:

$$\Gamma(\omega) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{+\infty} \gamma(\tau) e^{-j\omega\tau}$$

We can discard $\frac{1}{2\pi}$ since it is only a multiplicative factor.

Main properties

We highlight the single components of the spectrum's expression.

$$\Gamma(\omega) = ... + \gamma(-2)e^{+j2\omega} + \gamma(-1)e^{+j\omega} + \gamma(0) + \gamma(1)e^{-j\omega} + \gamma(2)e^{-j2\omega} + ...$$

Because the covariance function is an even function, therefore $\gamma(-1) = \gamma(+1)$, we have that:

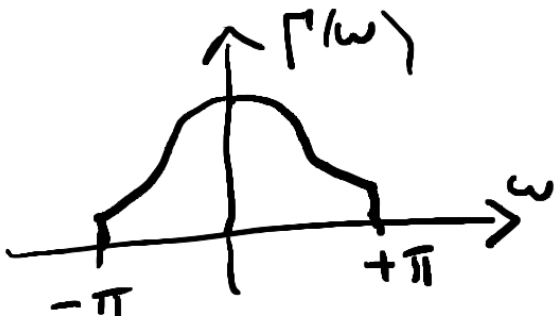
$$\gamma(-1)e^{+j\omega} + \gamma(1)e^{-j\omega} = \gamma(1)[e^{+j\omega} + e^{-j\omega}] = 2\gamma(1)cos(\omega)$$

By applying this operator to all the elements of the spectrum we obtain:

$$\Gamma(\omega) = \gamma(0) + 2\gamma(1)cos(\omega) + 2\gamma(2)cos(2\omega) + ...$$

Which is:

- a real function of the real variable ω.
- an even function
- a periodic function with period $T = 2\pi$
- $\Gamma(\omega) \geq 0$



Note that $\omega = 2\pi f$, and so $\omega = \pi \iff f = 0.5$, which is the maximum frequency in discrete time.
We can write the Fourier's antitransformation.

$$\gamma(\tau) = \mathscr{F}^{-1}[\Gamma(\omega)] = \frac{1}{2\pi}\int_{-\pi}^{+\pi}\Gamma(\omega)e^{j\omega\tau}\mathrm{d}\omega$$

In particular

$$\gamma(0) = \frac{1}{2\pi}\int_{-\pi}^{+\pi}\Gamma(\omega)\mathrm{d}\omega$$

Example

Let's consider the case of a white noise, with covariance $\gamma(\tau) = \begin{cases}\lambda^2, & \tau = 0\\ 0, & \tau \neq 0\end{cases}$

The spectrum is

$$\Gamma(\omega) = \gamma(0) + 2\gamma(1)cos(\omega) + 2\gamma(2)cos(2\omega) + ... = \gamma(0) = \lambda^2 \quad \text{(constant)}$$

Example
Let's consider the case of a MA(1) process.

$$v(t) = c_0\eta(t) + c_1\eta(t-1) \qquad \text{with } \eta \sim WN(0,\lambda^2)$$

We have:

$$\gamma(\tau) = \begin{cases}(c_0^2 + c_1^2)\lambda^2, & \tau = 0\\ c_0c_1\lambda^2, & \tau = \pm 1\\ 0, & \tau = \pm k, |k| > 1\end{cases}$$

Then we can express the spectrum as:

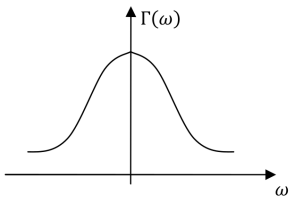
$$\Gamma(\omega) = [(c_0^2 + c_1^2 + 2c_0c_1cos(\omega)]\lambda^2$$

For $\omega = 0$ we have:

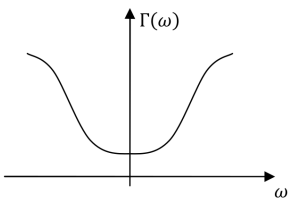
$$\Gamma(0) = [c_0^2 + c_1^2 + 2c_0c_1]\lambda^2 = (c_0 + c_1)^2\lambda^2$$

While for $\omega = \pi$ we have:

$$\Gamma(\pi) = [c_0^2 + c_1^2 - 2c_0c_1]\lambda^2 = (c_0 - c_1)^2\lambda^2$$



(a) Low frequency



(b) High frequency

Sometimes, for example for some ARMA processes, it is very hard to compute the covariance function. In this case, instead of computing the covariance and, from that, the spectrum, we can apply the so called "magic formula", which will be shown later.

4.2 Fundamental theorem of the spectral analysis

Given an ARMA process with a transfer function $W(z)$, an input $\eta(t) \sim WN(0,\lambda^2)$ and an output $v(t)$. The spectrum can be computed as:

$$\Gamma(\omega) = |W(e^{j\omega})|^2\lambda^2 = W(e^{j\omega})W(e^{-j\omega})\lambda^2$$

As a consequence $\Gamma(\omega) \geq 0$.
We define the complex spectrum as

$$\Phi(z) = \sum_{\tau=-\infty}^{+\infty}\gamma(\tau)z^{-\tau}$$

Then, by applying the magic formula, it can be rewritten as follows:

$$\Phi(z) = W(z)W(z^{-1})\lambda^2$$

The real spectrum is equivalent to the complex spectrum evaluated for $z = e^{j\omega}$.

$$\Gamma(\omega) = \Phi(z)|_{z=e^{j\omega}}$$

Example
Let's consider the case of a MA(1) process with $c_0 = c_1 = 1$.

$$v(t) = \eta(t) + \eta(t-1)$$

We can compute $\Gamma(\omega)$ from:

1. The definition
2. The magic formula
3. A graphical procedure based on the magic formula

1. The definition

The covariance function is

$$\gamma(\tau) = \begin{cases} 2\lambda^2, & \tau = 0 \\ \lambda^2, & \tau = \pm 1 \\ 0, & |\tau| > 1 \end{cases}$$

The complex spectrum is

$$\Phi(z) = \gamma(0) + \gamma(1)z^{-1} + \gamma(-1)z^{+1} + 0 + 0 + ... = (2 + z + z^{-1})\lambda^2$$

We can finally obtain the spectrum

$$\Gamma(\omega) = \Phi(z)|_{z=e^{-j\omega}} = (2 + e^{-j\omega} + e^{j\omega})\lambda^2 = (2 + 2cos\omega)\lambda^2$$

2. The magic formula

The system's behavior is

$$v(t) = \eta(t) + \eta(t-1) = (1 - z^{-1})\eta(t)$$

Then the transfer function is

$$W(z) = 1 + z^{-1} = \frac{z+1}{z}$$

We can compute the complex spectrum, which is the same as in the previous point.

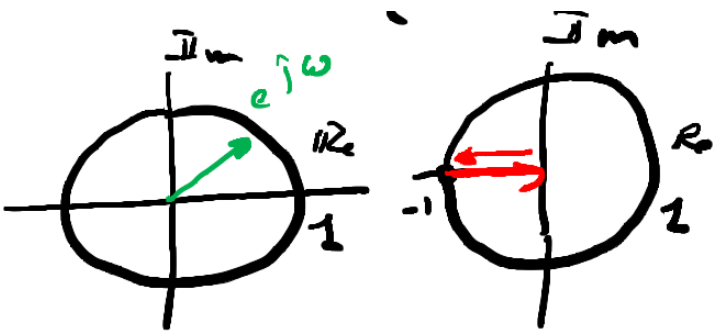
$$\Phi(z) = W(z)W(z^{-1})\lambda^2 = (1 + z^{-1})(1 + z)\lambda^2 = (2 + z + z^{-1})\lambda^2$$

3. Graphical study

We can compute the real spectrum from the expression of the transfer function

$$W(z) = \frac{z+1}{z} \quad \rightarrow \quad \Gamma(\omega) = |W(e^{j\omega})|^2\lambda^2 = \left|\frac{e^{j\omega}+1}{e^{j\omega}}\right|^2 \lambda^2$$

Then, graphically represent numerator and denominator

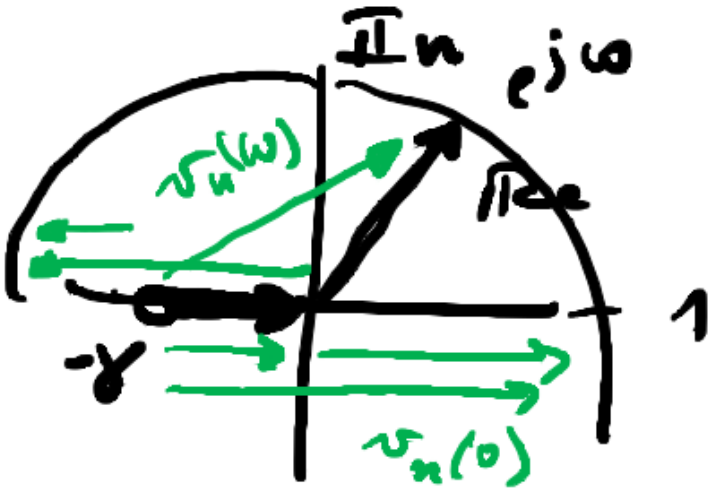


The denominator has modulus less than to 1 for each ω, while the numerator has modulus ranging from 0 for $\omega = \pi$ to 2 for $\omega = 0$.
Since the modulus of the transfer function is squared, the maximum value for $\Gamma(\omega)$ is equal to 4.
In general, suppose to have

$$W(z) = \beta \frac{z+\gamma}{z+\alpha}$$

which is an ARMA(1,1) process. We need to replace z with $e^{j\omega}$

$$W(e^{j\omega}) = \beta \frac{e^{j\omega}+\gamma}{e^{j\omega}+\alpha} = \beta \frac{v_n(\omega)}{v_d(\omega)}$$



We can draw γ as a vector starting from $-\gamma$ towards the origin, as shown in the previous figure. By summing it with the $e^{j\omega}$ term we obtain $v_n(\omega)$. When ω ranges from 0 to π, $v_n(0)$ starts from $-\gamma$ an goes horizontally towards the right up to the unit circle; $v_n(\pi)$ starts from $-\gamma$ and goes horizontally towards the left up to the unit circle. All the possible vectors in between starts from the zero $(-\gamma)$ and reach the unit circle.
The same reasoning can be applied for the poles of the denominator.
This allow us to identify that when a vector going from the origin of the disk to the border, when it reaches a zero, for that value of ω, the spectrum $|\Gamma(\omega)| = 0$; while when it reaches a pole, for that value of ω, the spectrum $|\Gamma(\omega)| = \infty$.

Note that from

$$\Phi(z) = W(z)W(z^{-1})\lambda^2$$

and

$$W(z) = \frac{\text{polynomial in } z}{\text{polynomial in } z}$$

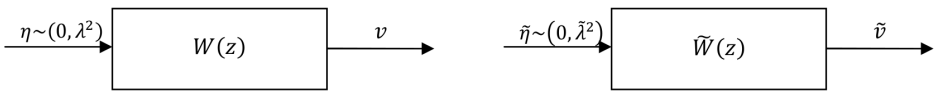
We have that

$$\Phi(z) = \frac{\text{polynomial in } z}{\text{polynomial in } z}\lambda^2$$

So, it is a rational function.

4.3 Multiplicity of ARMA models for a stationary process

Given a stochastic process, we want to find $W(z)$ from $\Phi(z)$. This problem is called spectral factorization problem. In this case, there are infinitely many different ARMA representations for the same process. Four examples can be mentioned:
1. Consider the following two systems



and $\tilde{W}(z) = z^{-1}W(z)$.
Then the complex spectrum for the second system is:

$$\tilde{\Phi}(z) = \tilde{W}(z)\tilde{W}(z^{-1})\tilde{\lambda}^2 = z^{-1}W(z)zW(z^{-1})\tilde{\lambda}^2$$

We can simplify the zs and by taking $\tilde{\lambda}^2 = \lambda^2$ we obtain:

$$\tilde{\Phi}(z) = W(z)W(z^{-1})\lambda^2 = \Phi(z)$$

It is possible to generalize to

$$\tilde{W}(z) = z^{-k}W(z) \quad \Rightarrow \quad \tilde{\Phi}(z) = \Phi(z)$$

2. By considering the same transfer functions, but linked by the relations

$$\tilde{W} = \frac{1}{\alpha}W \qquad \tilde{\lambda}^2 = \alpha^2\lambda^2$$

Then, we obtain:

$$\tilde{\Phi(z)} = \tilde{W(z)}\tilde{W}(z^{-1})\tilde{\lambda}^2 = \frac{1}{\alpha}W(z)\frac{1}{\alpha}W(z^{-1})\alpha^2\lambda^2 = W(z)W(z^{-1})\lambda^2 = \Phi(z)$$

3. This case is the simplification of the rescaling seen in the previous point

$$\tilde{W}(z) = \frac{z+\delta}{z+\delta}W(z)$$

It is trivial but sometimes the simplification may not be obvious. 4. Let's now consider a transfer function with reciprocal pole and zeros:

$$T(z) = \rho\frac{z+\alpha}{z+\frac{1}{\alpha}}$$

Hence, the value for $\Phi_{yy}(z)$ is

$$\Phi_{yy}(z) = T(z)T(z^{-1})\Phi_{uu}(z) = \rho\frac{z+\alpha}{z+\frac{1}{\alpha}}\rho\frac{z^{-1}+\alpha}{z^{-1}+\frac{1}{\alpha}}\Phi_{uu}(z) = \rho^2\frac{1+\alpha^2+\alpha z+\alpha z^{-1}}{1+\frac{1}{\alpha^2}+\frac{1}{\alpha}z+\frac{1}{\alpha}z^{-1}}\Phi_{uu}(z) = \alpha^2\rho^2\Phi_{uu}(z)$$

By taking $\rho^2 = \frac{1}{\alpha^2}$, the output spectrum coincides with the input spectrum. For this reason, this type of transfer function is called "all pass filter".

4.4 Canonical representation

Among the infinite representations of a transfer function, it is often useful to compute the canonical one, which is a representation satisfying the following conditions, that correspond to the inhibition of the four cases seen in the previous section:

1. Numerator and denominator have the same degree
2. Numerator and denominator are monic (the term with highest power has coefficient equal to 1).
3. Numerator and denominator are coprime
4. Numerator and denominator are stable polynomials: all poles and zeros of $W(z)$ are inside the unit disk

From a signal we can build the spectrum (or equivalently the covariance function). Once we have the spectrum, we can derive the canonical spectral factor and, thus, solve the prediction problem.
Given a rational process, there is one and only one ARMA representation which is canonical.

Important consequence

Given any transfer function with input u and output y, is it possible to invert the transfer function? The answer is: it depends.
For example with $W(z) = \frac{1}{z}$, the inverse is $W(z)^{-1} = z$ and while the output y of the direct system depends on the past of the input u, the output y, since the roles are inverted, of the inverse system depends on the future of the input u.
This happens because the degree of the denominator is greater than the degree of the numerator. But if we impose the numerator and denominator to have the same degree, the transfer function is invertible. The invertibility criteria requires also that the zeros and poles are inside the unit disk, so that also the inverse transfer function is stable (it is said to be stably invertible).

Chapter 5

Solving the prediction problem

As we said previously, the prediction problem consists of predicting the value $\hat{v}(t+r|t)$ of a signal $v(t+r)$ in a future time step $t+r$ from past values of $v(\cdot)$, i.e., $v(t), v(t-1), \ldots$. r is called prediction horizon. Suppose the signal is a stationary process as follows.

In the following sections, two problem will be solved:

1. Fake problem: we assume to know the past of $\eta(\cdot)$. Hence, $\hat{v}(t+r)$ will be computed given $\eta(t), \eta(t-1), \ldots$
2. True problem: $\hat{v}(t+r)$ given the past of $v(\cdot)$

5.1 The fake problem

By computing the long division between numerator and denominator of $\hat{W}(z)$ we obtain

$$\hat{W}(z) = \hat{w}_0 + \hat{w}_1 z^{-1} + \hat{w}_2 z^{-2} + \ldots$$

We then have

$$v(t+r) = \hat{W}(z)\eta(t+r) = \underbrace{\hat{w}_0\eta(t+r) + \hat{w}_1\eta(t+r-1) + \ldots + \hat{w}_{r-1}\eta(t+1)}_{\alpha(t)} + \underbrace{\hat{w}_r\eta(t) + \hat{w}_{r+1}\eta(t-1) + \ldots}_{\beta(t)}$$

Given that we know the past of $\eta(\cdot)$, then $\beta(t)$ can be computed.
Regarding $\alpha(t)$, we cannot know its values of $\eta(\cdot)$ because that would mean to know the future with respect to t. Furthermore, since η is a white noise, the knowledge of the past doesn't give any hint about possible future values. So, $\alpha(t)$ is fully unpredictable and the optimal fake predictor is

$$\hat{v}(t+r|t) = \beta(t) = \hat{w}_r\eta(t) + \hat{w}_{r+1}\eta(t-1) + \ldots$$

The prediction error is

$$v(t+r) - \hat{v}(t+r|t) = \hat{w}_0\eta(t+r) + \hat{w}_1\eta(t+r-1) + \ldots + \hat{w}_{r-1}\eta(t+1) = \alpha(t)$$

$$Var[\varepsilon(t)] = (\hat{w}_0^2 + \hat{w}_1^2 + \ldots + \hat{w}_{r-1}^2)\lambda^2$$

Note that the variance of the error increases with r.

5.1.1 Practical determination of the predictor

$\hat{v}(t+r|t) = \hat{w}_r\eta(t) + \hat{w}_{r+1}\eta(t-1) + \hat{w}_{r+2}\eta(t-2) + ... = (\hat{w}_r + \hat{w}_{r+1}z^{-1} + \hat{w}_{r+2}z^{-2} + ...)\eta(t) = \hat{W}_r(z)\eta(t)$

Starting from the result of the long division we have:

$$\begin{aligned}\hat{W}(z) &= \hat{w}_0 + \hat{w}_1 z^{-1} + ... + \hat{w}_{r-1}z^{-r+1} + \hat{w}_r z^{-r} + \hat{w}_{r+1}z^{-r-1} + ... \\ &= \hat{w}_0 + \hat{w}_1 z^{-1} + ... + \hat{w}_{r-1}z^{-r+1} + z^{-r}(\hat{w}_r + \hat{w}_{r+1}z^{-1} + \hat{w}_{r+2}z^{-2} + ...) \\ &= \hat{w}_0 + \hat{w}_1 z^{-1} + ... + \hat{w}_{r-1}z^{-r+1} + z^{-r}\hat{W}_r(z)\end{aligned}$$

Example
Consider the following an AR(1) process:

$$v(t) = av(t-1) + \eta(t)$$

the transfer function is a canonical factor:

$$\hat{W}(z) = \frac{z}{z-a}$$

To compute the 1-step predictor we first perform the long division

$$\begin{array}{ll|l} z & & z-a \\ \hline z & -a & 1 \\ \hline / & a & \end{array} \tag{5.1}$$

We then obtain

$$W(z) = 1 + \frac{a}{z-a} = 1 + z^{-1}\frac{az}{z-a} \quad \Rightarrow \quad \hat{W}_1(z) = \frac{az}{z-a}$$

The 2-step predictor is computed as follows:

$$\begin{array}{lll|l} z & & & z-a \\ \hline z & -a & & 1+az^{-1} \\ \hline / & a & & \\ & a & -a^2z^{-1} & \\ \hline & / & a^2z^{-1} & \end{array} \tag{5.2}$$

We then obtain

$$W(z) = 1 + az^{-1} + z^{-2}\frac{a^2 z}{z-a} \quad \Rightarrow \quad \hat{W}_2(z) = \frac{a^2 z}{z-a}$$

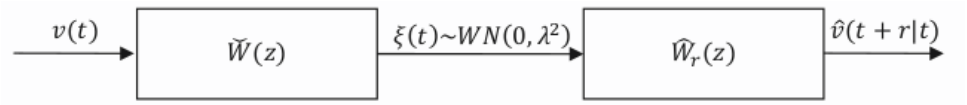
By generalization

$$\hat{W}_r(z) = \frac{a^r z}{z-a}$$

$$Var[\varepsilon(t)] = \begin{cases} 1^2\lambda^2, & \text{for } r=1 \\ (1^2 + a^2)\lambda^2, & \text{for } r=2 \\ (1^2 + a^2 + a^4)\lambda^2, & \text{for } r=3 \\ ... \end{cases}$$

5.2 The true problem

In the true problem we want to find the predictor from the data, i.e., from the past values of $v(\cdot)$. Suppose we have the following system, with the first block called whitening filter, and the second being the optimal fake predictor. The combination of the two transfer function will be the transfer function of



the oprimal r-step ahead predictor from the data.

$$W_r(z) = \check{W}(z)\hat{W}_r(z)$$

Suppose having the canonical spectral factor

$$\hat{W}(z) = \frac{C(z)}{A(z)} \quad \Rightarrow \quad \check{W}(z) = \frac{A(z)}{C(z)}$$

While the fake optimal predictor is obtained from the long division

$$\hat{W}_r(z)\frac{\ldots}{A(z)}$$

Thus, we the optimal predictor from data is

$$W_r(z) = \frac{A(z)}{C(z)} \cdot \frac{\ldots}{A(z)} = \frac{\ldots}{C(z)}$$

The optimal predictor from data is obtained from the optimal fake predictor by replacing its denominator with the numerator of the canonical spectral form.

Example
The usual AR(1) process:

$$v(t) = av(t-1) + \eta(t)$$

The optimal 1-step ahead predictor is

$$\hat{W}_1(z) = \frac{az}{z-a} \qquad \hat{W}(z) = \frac{z}{z-a} \quad \Rightarrow \quad W_1(z) = \frac{az}{z} = a$$

Hence

$$\hat{v}(t+1|t) = av(t)$$

Coming from a process

$$v(t+1) = av(t) + \underbrace{\eta(t+1)}_{\text{unpredictable}}$$

The optimal 2-steps ahead predictor is

$$\hat{W}_2(z) = \frac{a^2 z}{z-a} \qquad \frac{z}{z-a} \quad \Rightarrow \quad W_2(z) = \frac{a^2 z}{z} = a^2$$

Coming from a process

$$v(t+2) = av(t+1) + \eta(t+2) = a(av(t) + \eta(t+1)) + \eta(t+2) = a^2 v(t) + \underbrace{a\eta(t+1) + \eta(t+2)}_{\text{unpredictable}}$$

By generalization

$$\hat{v}(t+r|t) = a^r v(t)$$

Example
Let's consider the ARMA(1,1) process

$$v(t) = av(t-1) + \eta(t) + c\eta(t-1)$$

Note that the coefficient of $\eta(t)$ is directly 1 for simplicity, sooner or later, the coefficient would have been forced to that value to obtain a canonical form.
The process in operator form is:

$$A(z) = 1 - az^{-1} \qquad C(z) = 1 + cz^{-1} \qquad A(z)v(t) = C(z)\eta(t)$$

The process transfer function is

$$W(z) = \frac{C(z)}{A(z)} = \frac{1 + cz^{-1}}{1 - az^{-1}} = \frac{z + c}{z + a}$$

Let's apply the long division

$$\begin{array}{r|l} C(z) & A(z) \\ \hline A(z) & 1 \\ \hline C(z)-A(z) & \end{array} \tag{5.3}$$

$$W(z) = \frac{C(z)}{A(z)} = 1 + z^{-1} \underbrace{\frac{C(z)-A(z)}{A(z)} z}_{\substack{\text{optimal 1-step ahead} \\ \text{fake predictor } \hat{W}_1(z)}}$$

The optimal 1-step ahead predictor from data is

$$W_1(z) = \frac{C(z)-A(z)}{C(z)} z$$

By going back to the time domain

$$\hat{v}(t+1|t) = W_1(z)v(t) = \frac{C(z)-A(z)}{C(z)} z v(t)$$

We move $C(z)$ to the left-hand side of the equation

$$C(z)\hat{v}(t+1|t) = (C(z)-A(z))zv(t) = (C(z)-A(z))v(t+1)$$

The last equation is apparently contradicting the initial assumptions, because we are relying on $v(t+1)$ to predict $\hat{v}(t+1|t)$. But since in this case $C(z)$ and $A(z)$ are monic, by replacing them with their values we have

$$(1+cz^{-1})\hat{v}(t+1|t) = (1+cz^{-1}-1+az^{-1})v(t+1|t) = (a+c)z^{-1}v(t+1|t) = (a+c)v(t)$$

Finally, the expression of the estimate is

$$\hat{v}(t+1|t) = -c\hat{v}(t|t-1) + (a+c)v(t)$$

The variance of the prediction error is

$$Var[v(t+1) - \hat{v}(t+1|t)] = \lambda^2$$

In general, given the canonical form $\hat{W}(z) = 1 + z^{-1}\frac{C(z)-A(z)}{A(z)}z$, with $A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - ...$ and $C(z) = 1 + c_1 z^{-1} + c_2 z^{-2}$, from the equation

$$C(z)\hat{v}(t+1|t) = (C(z)-A(z))v(t+1)$$

we can derive a general solution

$$\hat{v}(t+1|t) = -c_1\hat{v}(t|t-1) - c_2\hat{v}(t-1|t-2) - ... + (a_1+c_1)v(t) + (a_2+c_2)v(t-1) + ...$$

Shortcut for predictor computation

The equation for computing the predictor can be obtained with a simpler process.
If we add and subtract $C(z)v(t)$ from the definition of the transfer function

$$A(z)v(t) \pm C(z)v(t) = C(z)\eta(t)$$

then we can divide by $C(z)$ and obtain the predictor by removing the white noise term, since it is completely unpredictable

$$v(t) = \frac{C(z)-A(z)}{C(z)}v(t) + \eta(t) \qquad \Rightarrow \qquad v(t|t-1) = \frac{C(z)-A(z)}{C(z)}v(t)$$

5.3 Prediction with exogenous signals

Let's suppose that the signal we want to predict depends also on another input variable $u(t)$, called exogenous variable. Differently from η, it is a deterministic variable.

Example (ARX)
In case of constant exogenous variable we have

$$v(t) = av(t-1) + u + \eta(t), \qquad \eta(t) \sim WN(0, \lambda^2) \qquad \Rightarrow \qquad u + \eta(t) \sim WN(u, \lambda^2)$$

With mean value

$$m = E[v(t)] = am + u + 0 \qquad \Rightarrow \qquad m = \frac{u}{1-a}$$

We can define a new process $\tilde{v}(t) = v(t) - m$ (debiased). Then we have

$$\tilde{v}(t) + m = a(\tilde{v}(t-1) + m) + u + \eta(t)$$

We can simplify some terms

$$\tilde{v}(t) = a\tilde{v}(t-1) + am - m + u + \eta(t) = a\tilde{v}(t-1) + \eta(t)$$

The predictor for the debiased process is obtained by the usual removal of the white noise term

$$\hat{\tilde{v}}(t|t-1) = a\tilde{v}(t-1)$$

By substitution we find the expression of the predictor of the original process

$$\hat{v}(t|t-1) = \hat{\tilde{v}}(t|t-1) + m = av(t-1) + am + m = av(t-1) + u$$

We can generalize the problem for a 1-step predictor for

$$A(z)v(t) = C(z)\eta(t) + B(z)u(t)$$

is given by

$$C(z)\hat{v}(t|t-1) = (C(z) - A(z))v(t) + B(z)u(t-1)$$

5.3.1 ARX process

In the same way an AR model of order n is defined as

$$v(t) = a_1 v(t-1) + a_2 v(t-2) + ... + a_n v(t-n) + \eta(t)$$

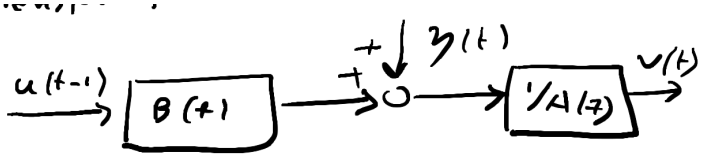
An ARX model of orders n_a, n_b is defined as

$$v(t) = a_1 v(t-1) + ... + a_{n_a} v(t-n_a) + b_1 u(t-1) + ... + b_{n_b} u(t-n_b) + \eta(t)$$

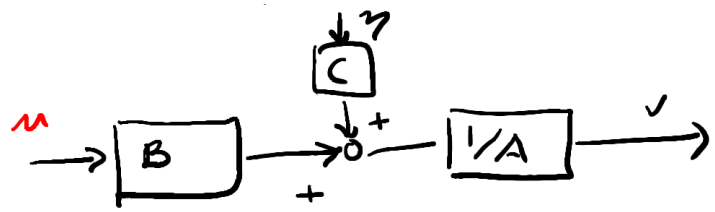
In operator form

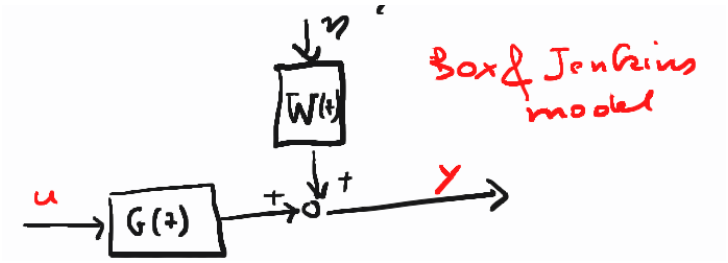
$$A(z)v(t) = B(z)u(t-1) + \eta(t)$$

With the transfer function from $u(t-1)$ to $v(t)$ equal to $\frac{B(z)}{A(z)}$ and the one from $\eta(t)$ to $v(t)$ equal to



$\frac{1}{A(z)}$.





5.3.2 ARMAX process

It is a process defined as

$$A(z)v(t) = C(z)\eta(t) + B(z)u(t-1), \qquad \eta \sim WN(0, \lambda^2)$$

The system can be represented with the Box&Jenkins model in which the white noise is considered as a disturb and $G(z)$ is the effect of the exogenous variable

$$y = G(z)u(t) + W(z)\eta(t)$$

So if the ARMAX process is described as

$$A(z)y(t) = B(z)u(t-1) + C(z)\eta(t)$$

By summing a subtracting $C(z)y(t)$ we obtain the respective predictor

$$C(z)y(t) = (C(z) - A(z))y(t) + B(z)u(t-1) + C(z)\eta(t)$$

We divide both sides by $C(z)$

$$y(t) = \underbrace{\frac{C(z) - A(z)}{C(z)}y(t) + \frac{B(z)}{C(z)}u(t-1)}_{\text{computable from the past values of } y \text{ and } u} + \eta(t)$$

The predictor is

$$\hat{y}(t|t-1) = \frac{C(z) - A(z)}{C(z)}y(t) + \frac{B(z)}{C(z)}u(t-1)$$

Part II

Identification

Chapter 6

Prediction Error Minimization (PEM) methods

The identification problem consists of estimating a model from the data. Assume having a system with input $u(\cdot)$ and output $y(\cdot)$, we would like to work out a model from their measurements.
Given a model, we can compute its output and compare it with the real output, to determine the prediction error:

$$\varepsilon(t) = y(t) - \hat{y}(t|t-1)$$

The goal is to obtain a prediction error which is both minimum and a white noise. The latter condition means that we cannot further improve the model because the error is completely unpredictable. The usual steps of the identification process are the following:

1. **Data collection** of the two series of data, $u(1), u(2), ..., u(N)$ and $y(1), y(2), ..., y(N)$.

2. **Choice of the models family**, represented as $\{M(\theta)|\theta \in \Theta\}$, where θ is a vector of parameters. We usually have AR and ARMA for time series; ARX and ARMAX for systems.

3. **Choice of the optimization criterion**: after computing the family of models in prediction form $\hat{M(\theta)}$ and its resulting prediction error, we choose the optimization criterion, like the mean squared error

   $$J(\theta) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon_\theta(t)^2$$

   where $\varepsilon_\theta(t)$ is the prediction error of model $M(\theta)$. Other criteria are possible, e.g., the mean absolute error

   $$J(\theta) = \frac{1}{N}\sum_{t=1}^{N} |\varepsilon_\theta(t)|$$

4. **Optimization**, in which the minimization is performed and the values of the model parameters are obtained

   $$\theta = \min J(\theta) = \min \frac{1}{N}\sum_{t=1}^{N} \varepsilon_\theta(t)^2$$

5. **Validation**: we need to perform a final analysis of the results, to evaluate if they satisfy our requirements. In the case of negative results, the choice of a new family of model may be necessary and the identification process has to be conducted again.

6.1 Least Squares method

It is the simplest method and it considers, as a family, all the ARX (or AR) models:

$$M(\theta) : y(t) = a_1 y(t-1) + ...a_{n_a} y(t-n_a) + b_1 u(t-1) + ... + u_{n_b} u(t-n_b) + \eta(t) = \theta' \varphi(t) + \eta(t)$$

Where θ is the parameter vector and $\varphi(t)$ is the observations vector:

$$\theta = \begin{vmatrix} a_1 \\ \vdots \\ a_{n_a} \\ b_1 \\ \vdots \\ b_{n_b} \end{vmatrix} \qquad \varphi(t) = \begin{vmatrix} y(t-1) \\ \vdots \\ y(t-n_a) \\ u(t-1) \\ \vdots \\ u(t-n_b) \end{vmatrix}$$

We can construct the prediction form of our family of models by simply removing the white noise term:

$$\hat{M}(\theta) : \hat{y}(t) = a_1 y(t-1) + ...a_{n_a} y(t-n_a) + b_1 u(t-1) + ... + u_{n_b} u(t-n_b) = \theta' \varphi(t)$$

We can solve the minimization problem by finding the parameters θ for which $\frac{\partial J}{\partial \theta} = 0$.

$$\begin{aligned} \frac{\partial J(\theta)}{\partial \theta} &= -\frac{1}{N} \sum_{t=1}^{N} 2 \left(y(t) - \theta' \varphi(t) \right) \varphi(t)' \\ &= -\frac{2}{N} \left(\sum_{t=1}^{N} y(t) \varphi(t)' - \sum_{t=1}^{N} \theta' \varphi(t) \varphi(t)' \right) \end{aligned}$$

Let's impose the derivative equal to zero.

$$\sum_{t=1}^{N} y(t) \varphi(t)' = \sum_{t=1}^{N} \theta' \varphi(t) \varphi(t)'$$

Notice that by swapping the two sides of the equations we obtain the **normal equations** $Ax = b$

$$\sum_{t=1}^{N} \varphi(t) \varphi(t)' \theta = \sum_{t=1}^{N} y(t) \varphi(t)'$$

Finally we have our parameters estimate

$$\hat{\theta} = \left[\sum_{t=1}^{N} \varphi(t) \varphi(t)' \right]^{-1} \sum_{t=1}^{N} y(t) \varphi(t)'$$

We need to verify that the found point is actually a minimum, by checking if $\frac{\partial^2 J(\theta)}{\partial \theta^2}$ is positive.
Remember that $\frac{\partial^2 J(\theta)}{\partial \theta^2}$ is an $N \times N$ matrix with the (i, j) entry defined as

$$\frac{\partial J(\theta)}{\partial \theta_i \partial \theta_j} \qquad i = 1, ..., N; j = 1, ..., N$$

In our case we have

$$\frac{\partial^2 J(\theta)}{\partial \theta^2} = \frac{2}{N} \left(\sum_{t=1}^{N} \varphi(t) \varphi(t)' \right)$$

Note that the last matrix is positive-semidefinite.
If we compute the Taylor series of $J(\theta)$ around the solution $\hat{\theta}$ of the normal equations we have

$$J(\theta) = J(\hat{\theta}) + \left. \frac{\partial J}{\partial \theta} \right|_{\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2} (\theta - \hat{\theta})' \left. \frac{\partial^2 J}{\partial \theta^2} \right|_{\hat{\theta}} (\theta - \hat{\theta}) + ...$$

Since $J(\theta)$ is a quadratic function, all the terms corresponding to a derivative greater that 2 are equal to zero, and since we imposed the first derivative to be zero we obtain

$$J(\theta) = J(\hat{\theta}) + \frac{1}{2} (\theta - \hat{\theta})' \frac{\partial^2 J(\theta)}{\partial \theta^2} (\theta - \hat{\theta})$$

But since we have said that $\frac{\partial^2 J(\theta)}{\partial \theta^2}$ is positive semi-definite, we have two possible cases:

1. if the matrix is positive definite, $J(\theta)$ is a paraboloid with vertex in $\hat{\theta}$
2. if the matrix is positive semi-definite but not positive definite, we have infinite solutions, as shown on the right-hand side of Figure 6.1

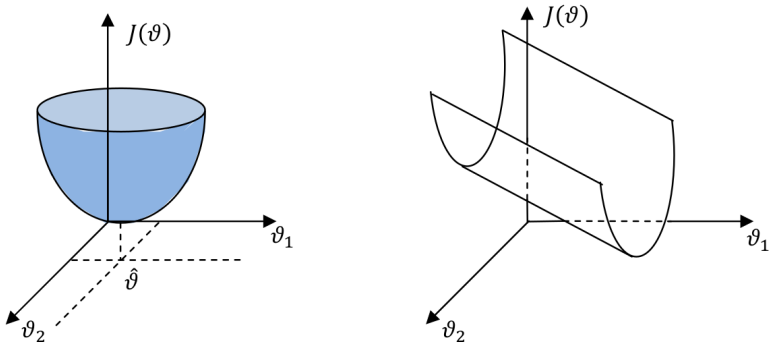


Figure 6.1: Possible shapes of $J(\theta)$: on the left-hand side if the matrix is positive definite, on the right-and side if the matrix is positive semi-definite but not positive definite

6.2 Identifiability

Now we can ask if the LS estimate is unique, which is called identifiability problem.
Let's consider the matrix $R(N)$ defined as:

$$R(N) = \frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi(t)'$$

If the matrix is positive semi-definite, it also invertible, then the normal equations have a unique solution.
In an ARX(1,1) process we have:

$$\varphi(t)\varphi(t)' = \begin{vmatrix} y(t-1) \\ u(t-1) \end{vmatrix} \begin{vmatrix} y(t-1) & u(t-1) \end{vmatrix} = \begin{vmatrix} y(t-1)^2 & y(t-1)u(t-1) \\ u(t-1)y(t-1) & u(t-1)^2 \end{vmatrix}$$

Hence, $R(N)$ can be defined as:

$$R(N) = \begin{vmatrix} \frac{1}{N}\sum y(t-1)^2 & \frac{1}{N}\sum y(t-1)u(t-1) \\ \frac{1}{N}\sum u(t-1)y(t-1) & \frac{1}{N}\sum u(t-1)^2 \end{vmatrix}$$

Note that the two elements on the diagonal are respectively the sample variance of y and the sample variance of u.
If we bring $N \to \infty$ we can derive $\bar{R}$:

$$\bar{R} = \begin{vmatrix} \bar{R}_{yy} & \bar{R}_{yu} \\ \bar{R}_{uy} & \bar{R}_{uu} \end{vmatrix} = \begin{vmatrix} \gamma_{yy}(0) & \gamma_{yu}(0) \\ \gamma_{uy}(0) & \gamma_{uu}(0) \end{vmatrix}$$

We can generalize the problem to a generic ARX model with

$$\varphi(t) = \begin{vmatrix} y(t-1) \\ \vdots \\ y(t-n_a) \\ u(t-1) \\ \vdots \\ u(t-n_b) \end{vmatrix}$$

Then, $\bar{R}_{uu}$ is a Toeplitz matrix:

$$\bar{R}_{uu} = \begin{vmatrix} \gamma_{uu}(0) & \gamma_{uu}(1) & \gamma_{uu}(2) & \\ \gamma_{uu}(1) & \gamma_{uu}(0) & \ddots & \ddots \\ \gamma_{uu}(2) & \ddots & \ddots & \\ & \ddots & & \end{vmatrix}$$

Inside each diagonal (not only the main one), the elements are the same, and inside the main diagonal we have the variance of u.
$\bar{R}_{yy}$ is the same, with y instead of u.
Finally $\bar{R}$ is

$$\bar{R} = \begin{vmatrix} \bar{R}_{yy} & \cdots \\ \cdots & \bar{R}_{uu} \end{vmatrix}$$

Thus, a necessary condition for the invertibility of $\bar{R}$ is that $\bar{R}_{uu}$ is invertible, in which case, u is said to be persistently exciting.

Example
In the case of $u(\cdot) \sim WN(0, \lambda^2)$ then, we have:

$$\bar{R}_{uu} = \begin{vmatrix} \gamma_{uu}(0) & \gamma_{uu}(1) & \gamma_{uu}(2) & \cdots \\ \gamma_{uu}(1) & \gamma_{uu}(0) & \gamma_{uu}(1) & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{vmatrix} = \begin{vmatrix} \lambda^2 & 0 & 0 & \cdots \\ 0 & \lambda^2 & 0 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{vmatrix} = \lambda^2 I$$

$\lambda^2 I$ is always invertible, so, the signal is persistently exciting for every possible order of the model.

Example
We have a system S described by the true transfer function

$$G^0(z) = \frac{z}{(z+0.5)(z+0.8)}$$

The generating mechanism of the data, i.e. the true transfer function, is unknown. Therefore, we can try by considering the following family of models

$$y(t) = a_1 y(t-1) + a_2 y(t-2) + a_3 y(t-3) + b_1 u(t-1) + b_2 u(t-2) + \eta(t)$$

and the vector of parameters is

$$\theta = \begin{vmatrix} a_1 \\ a_2 \\ a_3 \\ b_1 \\ b_2 \end{vmatrix}$$

Suppose, for example, that u is constant. We cannot identify the poles and zeros because the output would be constant as well. In fact, $\bar{R}_{uu}$ is not invertible, meaning that there are infinite many models providing the same estimate.
However, even if u is not constant, we have chosen a model which is oversized with respect to the true transfer function.

6.3 Estimation of mean, covariance and spectrum

The true generating mechanism $y(t)$ is usually not known, but we have the data $y(1), y(2), ..., y(N)$ collected from its realization.
For the same reason we don't know the real value of the mean, the covariance and the spectrum, but we can compute their estimated values from the data.
The estimate can be computed either directly from the data, or by finding a suitable model which represents the generating mechanism through the identification process and then by estimating the spectral properties of that model.
In this section we will refer to a generic estimator from the N data (e.g. the sample mean or sample covariance) as $\hat{s}_N$.

Correctness

$\hat{s}_N$ is a correct estimator if

$$E[\hat{s}_N] = \bar{s}$$

That is, the expected value of the estimator is equal to the probabilistic mean to be estimated.

Consistency

$\hat{s}_N$ is a consistent estimator if

$$Var[\hat{s}_N] \to 0 \quad \text{as } N \to 0$$

That is, the estimate error variance tends to zero as the number of measured data tends to infinity.

6.3.1 Mean value

It is possible to prove that the sample estimate of the expected value described as

$$\hat{m}_N = \frac{1}{N}\sum_{i=1}^{N} y(i)$$

is correct:

$$E[\hat{m}_N] = E\left[\frac{1}{N}\sum_{i=1}^{N} y(i)\right] = \frac{1}{N}\sum_{i=1}^{N} E[y(i)] = \frac{N}{N}\bar{m} = \bar{m}$$

and consistent:

$$\begin{aligned}
Var[\hat{m}_N] &= E[(\hat{m}_N - \bar{m})^2] \\
&= \frac{1}{N^2} E\left[\sum_{i=1}^{N}(y(i)-\bar{m})\sum_{j=1}^{N}(y(j)-\bar{m})\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{\tau=i-N}^{i-1}\underbrace{E[(y(i)-\bar{m})(y(i-\tau)-\bar{m})]}_{\gamma(\tau)} \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{\tau=i-N}^{i-1}\gamma(\tau) \\
&= \frac{1}{N}\sum_{\tau=1-N}^{N-1}\frac{N-|\tau|}{N}\gamma(\tau) \le \frac{1}{N}\sum_{\tau=1-N}^{N-1}|\gamma(\tau)| \to 0 \quad \text{as } N \to \infty
\end{aligned}$$

6.3.2 Covariance

For sake of simplicity, we assume a zero mean process.
We want to consider sampled estimators which have the main properties of the covariance function, such as:

1. Positive, i.e. $\gamma(0) > 0$
2. Even, i.e $\gamma(\tau) = \gamma(-\tau)$
3. $\gamma(0) > |\gamma(\tau)| \quad \forall \tau \neq 0$
4. Has positive semi-definite Toeplitz matrix

Two different estimators are possible:

$$\hat{\gamma}_N^a(\tau) = \frac{1}{N}\sum_{i=1}^{N-|\tau|} y(i)y(i+|\tau|)$$

$$\hat{\gamma}_N^b(\tau) = \frac{1}{N-|\tau|}\sum_{i=1}^{N-|\tau|} y(i)y(i+|\tau|)$$

Properties 1, 2 and 3 are satisfied by both the estimators. Regarding property 4, a matrix is positive semidefinite if there exists a matrix T such that M=TT'.
If we define

$$T = \begin{bmatrix}
y(1) & y(2) & \cdots & y(N) & 0 & 0 & \cdots & 0 \\
0 & y(1) & \cdots & y(N-1) & y(N) & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & y(1) & y(2) & y(3) & \cdots & y(N)
\end{bmatrix}$$

Note that

$$T = \begin{bmatrix} \hat{\gamma}_N(0) & \hat{\gamma}_N(1) & \cdots & \hat{\gamma}_N(N-1) \\ \hat{\gamma}_N(1) & \hat{\gamma}_N(0) & \cdots & \hat{\gamma}_N(N-2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\gamma}_N(N-1) & \hat{\gamma}_N(N-2) & \cdots \hat{\gamma}_N(0) & \end{bmatrix} = \frac{1}{N} TT' \geq 0$$

Which is valid only for estimator a, hence estimator b does not satisfy property 4.
On the other hand the estimator a is not correct:

$$E[\hat{\gamma}_N(\tau)] = \frac{1}{N} E\left[\sum_{i=1}^{N-|\tau|} y(i)y(i+|\tau|)\right] = \frac{N-|\tau|}{N}\gamma(\tau) \neq \gamma(\tau)$$

However, it is asymptotically correct

$$E[\hat{\gamma}_N(\tau)] = \frac{N-|\tau|}{N}\gamma(\tau) \to \gamma(\tau) \quad \text{as } N \to +\infty$$

While estimator b is correct for all N:

$$E[\hat{\gamma}_N(\tau)] = \frac{1}{N-|\tau|} E\left[\sum_{i=1}^{N-|\tau|} y(i)y(i+|\tau|)\right] = \frac{N-|\tau|}{N-|\tau|}\gamma(\tau) = \gamma(\tau)$$

Both the estimators are consistent in case of stationary ARMA processes.

6.3.3 Spectrum

To obtain a sampled estimator of the spectrum we repeat infinitely many times the data series

$$..., y(N), y(1), y(2), ..., y(N), y(1), y(2), ..., y(N), y(1), y(2), ...$$

This series is a periodic signal, which can thus be written as the antitransform of a signal a_k:

$$y(t) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} a_k e^{j\omega_k t} \quad \text{where } \omega_k = 2\pi\frac{K}{N}$$

In other words, the periodic extension can be seen as the sum of N sinusoids, with periodicity given by $\omega = \frac{2\pi}{N}, \frac{2\pi}{N}2, \cdots, 2\pi$ and periods $T = \frac{2\pi}{\omega} = N, \frac{N}{2}, \cdots, 1$. The amplitude of such armonic components is

$$a_k = \frac{1}{\sqrt{N}} \sum_{t=1}^{N} y(t) e^{-j\omega_k t}$$

In particular, $|a_k|^2$ is the power of each of these armonic components of the signal. This defines the periodogram:

$$\begin{aligned} \hat{\Gamma}(\omega_k) = |a_k|^2 &= \frac{1}{N} \sum_{t,s=1}^{N} y(t)y(s)^* e^{-j\omega_k t} e^{j\omega_k s} \\ &= \frac{1}{N} \sum_{t,s=1}^{N} y(t)y(s)^* e^{-j\omega_k (t-s)} \\ &= \sum_{\tau=1-N}^{N-1} \underbrace{\left(\frac{1}{N}\sum_{t=1}^{N} y(t)y(t-\tau)^*\right)}_{\text{covariance estimator of type } a} e^{-j\omega_k \tau} \\ &= \sum_{\tau=1-N}^{N-1} \hat{\gamma}_N(\tau) e^{-j\omega_k t} \end{aligned}$$

Note that the expression of $\hat{\Gamma}$ is very similar to the definition of the spectrum.
As an estimator, it is asymptotically correct:

$$E[\hat{\Gamma}(\omega)] \to \Gamma(\omega) \text{ as } N \to +\infty$$

and not consistent

$$Var[\hat{\Gamma}(\omega)] \to \Gamma(\omega)^2 \text{ as } N \to +\infty$$

The problem of the inconsistency can be tackled using the Bartlett method.

6.3.4 Bartlett method

Given N data (with N "large"), we divide the data series in r non-overlapping sub-series of length $\hat{N} = \frac{N}{r}$ ($N >> r$).
We then obtain one periodogram for each sub-series

$$\hat{\Gamma}^{(i)}_{\hat{N}}(\omega), \quad i, ..., r$$

We compute the average periodogram:

$$\bar{\hat{\Gamma}}(\omega) = \frac{1}{r} \sum_{i=1}^{r} \hat{\Gamma}^{(i)}_{\hat{N}}(\omega)$$

Under the assumption that the data of different sub-series are uncorrelated between each other (that's why $N >> r$ is needed), then

$$Var[\bar{\hat{\Gamma}}_{\hat{N}}(\omega)] \simeq \frac{1}{r}\Gamma^2(\omega)$$

The uncertainty is now significantly reduced.

6.4 Gain of a dynamic system

Let's consider a generic dynamic system described by the transfer function

$$G(z) = \frac{N(z)}{D(z)}$$

If $u(t) = \bar{u}$ is constant, and the system is stable, we expect $y(t) = \bar{y}$ to be constant as well. Then, the value

$$\mu = \frac{\bar{y}}{\bar{u}}$$

is the gain of the system, and we would like to compute it from $G(z)$.
We have

$$y(t) = G(z)u(t) = \frac{N(z)}{D(z)}u(t)$$

We can move the denominator of the transfer function to the left-hand side:

$$D(z)y(t) = N(z)u(t)$$

Suppose $D(z)$ to be:

$$D(z) = d_0 z^n + d_1 z^{n-1} + ...$$

Then,

$$D(z)y(t) = d_0 y(t+n) + d_1 y(t+n-1) + ...$$

If the output is constant:

$$D(z)y(t) = d_0\bar{y} + d_1\bar{y} + ... = (d_0 + d_1 + ...)\bar{y} = D(z)|_{z=1}\,\bar{y}$$

In the same way:

$$N(z)u(t) = N(z)|_{z=1}\,\bar{u}$$

We can combine the two expressions as:

$$D(z)|_{z=1}\,\bar{y} = N(z)|_{z=1}\,\bar{u}$$

The gain can be finally computed as the transfer function evaluate $z = 1$:

$$\mu = \frac{\bar{y}}{\bar{u}} = \left.\frac{N(z)}{D(z)}\right|_{z=1}$$

Example
Assume that the data are generated by

$$G^0(z) = \frac{z}{(z+0.5)(z+0.8)} = \frac{z}{z^2+1.3z+0.4}$$

the true gain is:

$$\mu^0 = \frac{1}{1^2+1.3\cdot 1+0.4} = \frac{1}{2.7}$$

There is only one zero at $z = 0$ and two poles at $z = -0.5$ and $z = -0.8$, hence, the system is stable.
We want to identify this system and we write it as an ARX model:

$$M(\theta) : a_1 y(t-1) + a_2 y(t-2) + b_1 u(t-1) + \eta(t)$$

The transfer function of M is

$$G(z) = \frac{b_1 z^{-1}}{1 - a_1 z^{-1} - a_2 z^{-2}} = \frac{b_1 z}{z^2 - a_1 z - a_2}$$

Note that $G(z) = G^0(z)$ if $b_1 = b_1^0 = 1$, $a_1 = a_1^0 = 1.3$ and $a_2 = a_2^0 = 0.4$. We need to estimate the value of these parameters from the data.
Starting from the measurements at times $t = 1, 2, ..., N$, we can find θ by solving the normal equations.
Let's consider two cases:

Case A

$u(t) = \bar{u}$ is constant and thus the output $y(t) = \bar{y}$ is constant, too. The observation vector is

$$\varphi(t) = \begin{vmatrix} \bar{y} \\ \bar{y} \\ \bar{u} \end{vmatrix}$$

The $\bar{R} = \sum \varphi(t)\varphi(t)'$ matrix is singular (i.e. not invertible).
The only information we can obtain is the gain of the system

$$\mu = \frac{\bar{y}}{\bar{u}} = \frac{b_1}{1 - a_1 - a_2}$$

In fact, there are infinitely may combinations of parameters that give as a result the observed gain.

Case B

$u(t) \sim WN(0, \lambda^2)$ is a white noise and thus $y(t)$ is a stochastic process. The parameter estimation with N samples is

$$\hat{\theta}_N = \left(\sum_{t=1}^{N} \varphi(t)\varphi(t)'\right)^{-1} \sum_{t=1}^{N} \varphi(t) y(t)$$

We have that for $N \to \infty$

$$\hat{\theta}_N \to \theta^0$$

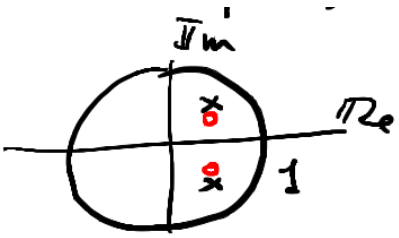
Let's now consider another possible model class

$$M : y(t) = a_1 y(t-1) + a_2 y(t-2) + a_3 y(t-3) + b_1 u(t-1) + b_2 u(t-2)$$

In this case, the parameters are not uniquely identifiable even if u is a white noise, because the model class is oversized. The model transfer function has the following structure:

$$G(z)\frac{z(b_1 z + b_2)}{(\quad)(\quad)(\quad)}$$

The denominator has degree equal to 3, while that true transfer function's one has degree 2. In order to have $G(z) = G^0(z)$ we need to perform a simplification between the numerator and a factor of the denominator. The number of these simplifications is infinite, so infinitely many equivalent models are possible.
A possible way to identify cases in which the model is too complex is by inspecting the poles and zeros of the estimated transfer function and checking if there are poles very close in value to some zeros and by performing the simplification. We can summarize the concepts of identifiability seen so far by saying



that $\bar{R}$ is invertible if and only if the following conditions are satisfied:

1. **Experimental identifiability condition**: the input is persistently exciting with an order greater or equal to the number of parameters n_b, associated to u, to be estimated. It depends on the experiment performed.
2. **Structural identifiability condition**: there are no simplifications. It depends on the complexity of the adopted class of models.

6.5 Maximum Likelihood methods

This method differs from the Least Squares method because it is based on a different class of models, ARMA (or ARMAX) instead of AR (or ARX). This means that there is no more linearity in the parameters and no normal equations.
For the ARMAX family we have:

$$M : A(z)y(t) = B(z)a(t-1) + C(z)\eta(t)$$

where

$$A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - ...$$
$$C(z) = 1 + c_1 z^{-1} + c_2 z^{-2} + ...$$
$$B(z) = b_1 + b_2 z^{-1} + ...$$

The vector of parameters of M is

$$\theta' = |\underbrace{a_1 \quad a_2 \quad ... \quad a_{n_a}}_{A(z)} \quad \underbrace{b_1 \quad b_2 \quad ... \quad b_{n_b}}_{B(z)} \quad \underbrace{c_1 \quad c_2 \quad ... \quad c_{n_c}}_{C(z)}|$$

As in the previous case, we need to find a suitable $\hat{\theta}$ from the data, which is, as always, $y(1), y(2), ..., y(N)$, $u(1), u(2), ...u(N)$.
The performance index based on the prediction error still can be the mean squared error

$$J = \frac{1}{N}\sum_{t=1}^{N} \varepsilon_\theta(t)^2$$

Differently from the LS method, the function is now non-convex. Iterative methods, such as the Newton method, can be used to solve the minimization problem.

6.5.1 The Newton method

Let's suppose, without loss of generality, that θ is a scalar. The method is based on approximating J with a quadratic function $V(\theta)$. The minimum of this function for the r^{th} iteration is considered as the estimated vector of parameters $\hat{\theta}^{(r+1)}$ of the following iteration.
By letting $r \to \infty$ we obtain the minimum $\hat{\hat{\theta}}$. But there is no guarantee that the minimum found is a global minimum. One simple method to deal with this problem is to execute multiple times the algorithm with different initializations and take the best among the different runs.
If you consider a quadratic approximation, the approximating function can be obtained by the Taylor development:

$$V(\theta) = J(\theta)|_{\theta=\theta^{(r)}} + \left.\frac{\partial J(\theta)}{\partial \theta}\right|_{\theta=\theta^{(r)}} (\theta - \theta^{(r)}) + \frac{1}{2}(\theta - \theta^{(r)})' \left.\frac{\partial^2 J(\theta)}{\partial \theta^2}\right|_{\theta=\theta^{(r)}} (\theta - \theta^{(r)})$$

The minimum of this function is computed as follows (Newton formula):

$$\theta^{(r+1)} = \theta^{(r)} - \left(\left.\frac{\partial^2 J(\theta)}{\partial \theta^2}\right|_{\theta=\theta^{(r)}} \right)^{-1} \left.\frac{\partial J(\theta)}{\partial \theta}\right|'_{\theta=\theta^r}$$

The first and second order derivatives of the error with respect to the parameters, in the case of Mean Squared Error, are respectively:

$$\frac{\partial J(\theta)}{\partial \theta} = \frac{2}{N} \sum_{t=1}^{N} \varepsilon(t) \frac{\partial \varepsilon(t)}{\partial \theta}$$

$$\frac{\partial^2 J(\theta)}{\partial \theta^2} = \frac{2}{N} \sum_{t=1}^{N} \frac{\partial \varepsilon(t)}{\partial \theta} \frac{\partial \varepsilon(t)}{\partial \theta} + \frac{2}{N} \sum_{t=1}^{N} \varepsilon(t) \frac{\partial^2 \varepsilon(t)}{\partial \theta^2}$$

The second term is usually neglected for simplicity.
We can define the vector

$$\psi(t) = -\frac{\partial \varepsilon(t)}{\partial \theta}$$

By replacing the expressions of the derivatives into the Newton formula we obtain the Gauss-Newton formula:

$$\theta^{(r+1)} = \theta^{(r)} + \left(\sum_{t=1}^{N} \psi(t)\psi(t)' \right)^{-1} \sum_{t=1}^{N} \psi(t)\varepsilon(t)$$

If we change it a little, the formula resembles the normal equations:

$$\sum_{t=1}^{N} \psi(t)\psi(t)' \underbrace{\left(\theta^{(r+1)} - \theta^{(r)}\right)}_{\theta} = \sum_{t=1}^{N} \psi(t) \underbrace{\varepsilon(t)}_{y(t)}$$

Let's see how to compute $\varepsilon(\cdot)$ and $\psi(\cdot)$ from the data, by considering the following models family:

$$M : Ay(t) \pm Cy(t) = Bu(t-1) + C\eta(t)$$

From that, we obtain:

$$Cy(t) = [C - A]\, y(t) + Bu(t-1) + C\eta(t)$$

The output's expression is:

$$y(t) = \underbrace{\frac{C-A}{C} y(t)}_{\text{past of } y} + \underbrace{\frac{B}{C} u(t-1)}_{\text{past of } u} + \eta(t)$$

C and A are monic polynomials, so we have:

$$C = 1 + c_1 z^{-1} + c_2 z^{-2} + ...$$

$$A = 1 - a_1 z^{-1} - a_2 z^{-2}$$

Thus, $C - A$ is:

$$C - A = (a_1 + c_1) z^{-1} + (a_2 + c_2) z^{-2} + ...$$

Note that the known term is absent, so it is a function of the past time points.
The predictor is computed as usual by dropping the white noise term:

$$\hat{y}(t) = \frac{C-A}{C}y(t) + \frac{B}{C}u(t-1)$$

We multiply by C on both sides

$$C\hat{y}(t) = Cy(t) - Ay(t) + Bu(t-1)$$

We move the $Cy(t)$ term on the left-hand side:

$$C(\underbrace{y(t) - \hat{y}(t)}_{\varepsilon(t)}) = Ay(t) - Bu(t-1)$$

Finally the prediction error equation for iteration r is:

$$C^{(r)}\varepsilon(t)^{(r)} = A^{(r)}y(t) - B^{(r)}u(t-1)$$

The following steps outline the entire iterative process to find the estimate $\hat{\theta}$ of the parameters:

1. At iteration r we have the estimate $\hat{\theta}^{(r)}$ of the parameters
2. From $\theta^{(r)}$, obtain $A^{(r)}(z)$, $B^{(r)}(z)$ and $C^{(r)}(z)$
3. Filter the data with such polynomials to obtain $\varepsilon(t)^{(r)}$
4. Filter the data to obtain $\psi(t)^{(r)}$
5. Use Gauss-Newton formula to compute $\hat{\theta}^{(r+1)}$
6. Repeat until convergence

6.6 Performance of prediction error identification methods

If we construct the prediction error as usual as:

$$\varepsilon_\theta(t) = y(t) - \hat{y}_\theta(t)$$

Both $y(t)$ and $\hat{y}_\theta(t)$ are sequences of points. This means that the performance index depends on the specific points that are provided. To highlight it, we add the subscript N:

$$J_N(\theta) = \frac{1}{N}\sum_{t=1}^{N}\varepsilon_\theta(t)^2$$

Then also the estimated parameters $\hat{\theta}_N$ depends on the data points.
If the prediction error can be seen as a stationary process, then, under mild conditions, we expect that with $N \to \infty$

$$J_N(\theta) \to \bar{J}(\theta) = E[\varepsilon_\theta(t)^2]$$

Note that $\bar{J}(\theta)$ does not depend on the particular outcome of the random experiment.
If we assume θ to be a scalar for simplicity and we represent the minimization functions of $J_N(\theta)$ depending on the number of data points N, as shown in Figure 6.2

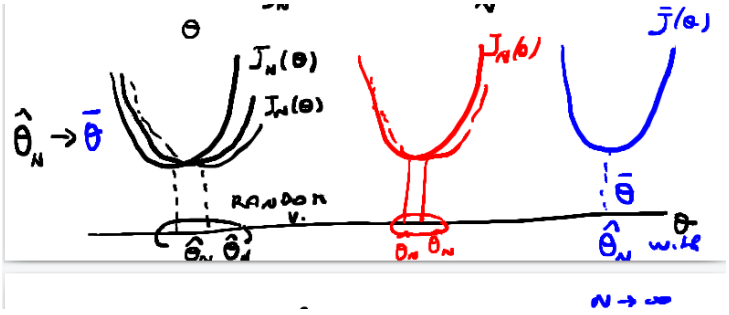


Figure 6.2: The set of minimum points for θ_N for three increasing values of N. The right-most one is for $N \to \infty$

The higher the N, the closer the minimum values of $\hat{\theta}_N$ to each other for different experiments (i.e. for different set of data points), up to when $N \to \infty$ in which there is only a single function $\bar{J}(\theta)$ and thus a single minimum value $\bar{\theta}$ that doesn't depend on the experiments outcome.
Since for finite N, $\hat{\theta}_N$ depends on the outcome of the experiment, it can be considered as a random variable, so we expect that the random variable $\hat{\theta}_N$ tends, for $N \to \infty$ to $\bar{\theta}$, where $\bar{\theta}$ is the result of the minimization process of the asymptotic performance index:

$$\bar{\theta} = \min E[\varepsilon_\theta(t)^2]$$

Example
Assume the data generator to be:

$$S : y(t) = a^0 y(t-1) + \eta(t), \qquad \text{with } \eta(\cdot) \sim WN(0, \lambda^2)$$

Where a_0 is a "true" parameter. Let's suppose $|a^0| < 1$ so that $y(\cdot)$ is a stationary process.
We consider the model:

$$M(\theta) : \hat{y}_\theta(t) = ay(t-1)$$

In this case the vector of parameters reduces to a vector with a single value $\theta = [a]$.
We estimate $\hat{a}_N$ with the Least Squares method. The vector of observations has size 1×1

$$\varphi(t) = |y(t-1)|$$

The normal equations can be written as

$$\sum_{t=1}^{N} y(t-1)^2 \theta = \sum_{t=1}^{N} y(t-1)y(t)$$

The parameters estimate is

$$\hat{\theta} = \hat{a} = \frac{\frac{1}{N}\sum_{t=1}^{N} y(t-1)y(t)}{\frac{1}{N}\sum_{t=1}^{N} y(t-1)^2} = \frac{\hat{\gamma}_{yy}(1)}{\hat{\gamma}_{yy}(0)}$$

Let's consider the asymptotic performance index

$$\bar{J}(\theta) = E[\varepsilon_\theta(t)^2]$$

If $\varepsilon_\theta(\cdot)$ is stationary, then $E[\varepsilon_\theta(t)^2]$ does not depend on t.
In this case the prediction error is:

$$\varepsilon_\theta(t) = y(t) - \hat{y}(t) = a^0 y(t-1) + \eta(t) - ay(t-1) = (a^0 - a)y(t-1) + \eta(t)$$

$$\varepsilon_\theta(t)^2 = (a^0 - a)^2 y(t-1) + \eta(t)^2 + 2(a^0 - a)y(t-1)\eta(t)$$

$$E[\varepsilon_\theta(t)^2] = (a^0 - a)^2 E[y(t-1)^2] + \lambda^2 + 2(a^0 - a)E[y(t-1)\eta(t)]$$

$y(t-1)$ is a function of $\eta(t-1), \eta(t-2), ...$ so it is uncorrelated with $\eta(t)$ and $E[y(t-1)\eta(t)] = 0$.
The performance index can be written as:

$$\bar{J}(\theta) = (a^0 - a)^2 \gamma_{yy}(0) + \lambda^2$$

For $N \to \infty$, $\hat{a}_N \to a^0$, until, for $a = a^0$ we have that $\bar{J}(\theta) = \lambda^2$.
In general, if the family of models contains the data generation mechanism S, i.e. $\exists \bar{\theta}$ such that $M(\bar{\theta}) = S$, then the minimum of the performance index is obtained for a $\bar{\theta}$ coinciding with the true parameter.
Note that there may be a multiplicity of points of minimum Δ. In that case $\bar{\theta}$ coincides with one of them. In this case, it may be useful to downgrade the complexity of the model to obtain a unique value.

Example
Let's solve the previous example, but with a slightly different computation.

$$\begin{aligned}
\bar{J}(\theta) &= E[\varepsilon_\theta(t)^2] \\
&= E[(y(t) - \underbrace{ay(t-1)}_{\hat{y}_\theta(t)})^2] \\
&= E[y(t)^2] + a^2 E[y(t-1)^2] - 2aE[y(t)y(t-1)] \\
&= \gamma_{yy}(0) + a^2 \gamma_{yy}(0) - 2a\gamma_{yy}(1)
\end{aligned}$$

We want to find the minimum with respect to a, so we compute its derivative:

$$\frac{\partial \bar{J}}{\partial a} = 2a\gamma_{yy}(0) - 2\gamma_{yy}(1)$$

By setting its value to zero, we derive the Yule-Walker equation:

$$a\gamma_{yy}(0) = \gamma_{yy}(1)$$

Finally, the same expression is obtained

$$\bar{a} = \frac{\gamma_{yy}(1)}{\gamma_{yy}(0)}$$

Example
Now consider as data generation mechanism a MA(1) process, as follows:

$$S : y(t) = \eta(t) + c^0\eta(t-1), \qquad \eta \sim WN(0, \lambda^2)$$

The family of model is the set of AR(1) models:

$$\hat{M} : \hat{y}(t) = ay(t-1)$$

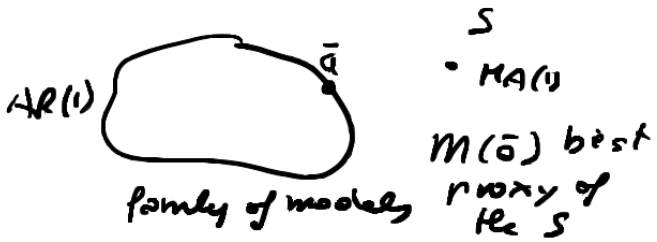
Again, the only parameter to be estimated is $\theta = a$.
The set of models does not include the data generation mechanism. It may happen since the data generation mechanism is not known.
Knowing that in MA(1):

$$\begin{aligned}\gamma_{yy}(0) &= (1^2 + c^{0\,2})\lambda^2 \\ \gamma_{yy}(1) &= 1 \cdot c^0 \cdot \lambda^2 \\ \gamma_{yy}(2) &= 0 \\ &\vdots\end{aligned}$$

As we have, seen, the parameter can be estimated as:

$$\bar{a} = \frac{\gamma_{yy}(1)}{\gamma_{yy}(0)} = \frac{c^{0\,2}}{1 + c^{0\,2}}$$

The estimated parameter is not the true one, but it is the best proxy model in the chosen family.



6.7 Validity test of the estimated model

We would like to check if the family of models does not include the data generation mechanism. It basically consists of detecting if the prediction error is a white noise.

Case A

$$\begin{aligned}S : AR(1) : y(t) &= a^0 y(t-1) + \eta(t) \\ M : AR(1) : \hat{y}(t) &= ay(t-1)\end{aligned}$$

with the prediction error identification method, the estimate $\hat{a}_N$ tends to a^0.

$$\hat{t}(t) \to = a^0 y(t-1) \quad \Rightarrow \quad \varepsilon(t) = \eta(t), \qquad \eta \sim WN$$

Case B

$$S : MA(1) : y(t) = \eta(t) + c^0\eta(t-1)$$
$$M : AR(1) : \hat{y}(t) = ay(t-1)$$

As seen before $\hat{a}_N \to \bar{a} = \frac{c^0}{1+c^{0\,2}}$.
Consequently,

$$\varepsilon(t) = y(t) - \hat{y}(t) = \eta(t) + c^0\eta(t-1) - \frac{c^0}{1+c^{0\,2}} \underbrace{y(t-1)}_{f(\eta(t-1),\eta(t-2))}$$

So its function of $\eta(t), \eta(t-1), \eta(t-2)$ and, hence, it is not a white noise.

Example
Assume an ARMAX data generation mechanism:

$$S : y(t) = a^0y(t-1) + b^0u(t-1) + \eta(t) + c^0\eta(t-1)$$

And an ARX family of models:

$$\hat{M} : \hat{y}(t) = ay(t-1) + bu(t-1)$$

The estimated parameters, for $N \to \infty$, are:

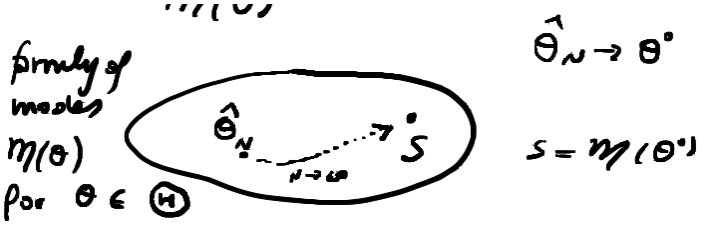
$$\bar{a} = a^0 + c^0\frac{Var[\eta]}{Var[y]}$$
$$\bar{b} = b^0$$

If $c^0 = 0$, so that S is in ARX, then $\bar{a} = a^0$.

6.8 Summary

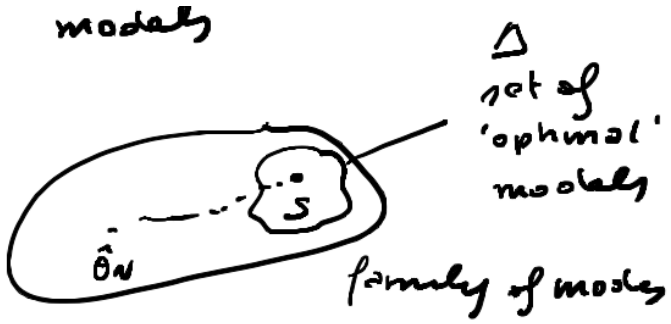
We can represent the possible situations depending on the data generation mechanism S and the family of models $M(\theta)$.

1. If $M(\theta)$ includes S, we have that $S = M(\theta^0)$



2. If $M(\theta)$ does not include S, the model $M(\bar{\theta})$ is the best proxy of S in the considered family of models

3. There is a multiplicity of optimal models Δ that includes S



4. There is a multiplicity of optimal models Δ that does not include S

6.9 Anderson Whiteness Test

If the model obtained through the identification process is actually the true model (i.e. $M(\hat{\theta}) = S$) then, the prediction error is a white noise.
The Anderson Whiteness Test allows to conclude (on a probabilistic basis) on the whiteness of the prediction error.
We define the normalized correlation function:

$$\hat{\rho} = \frac{\hat{\gamma}_N^a(\tau)}{\hat{\gamma}_N^a(0)} \quad \tau \geq 0$$

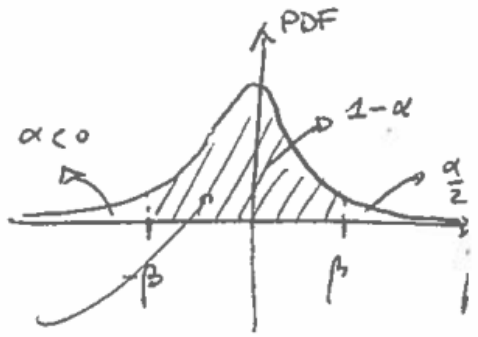
The test is based on the following assumption: if the process generating the prediction error $\varepsilon(t)$ is a white noise, then for $\tau > 0$ (and for $N >> 0$), $\hat{\rho}(\tau)$ has the following properties:

1. $\hat{\rho}(\tau) \sim \mathcal{N}(0, \frac{1}{N})$
2. $\hat{\rho}(i) \perp \hat{\rho}(j) \quad \forall i \neq 0$

We can therefore consider $\hat{\rho}(\tau), \tau = 1, 2, ...$ as independent values of a gaussian variable.
So, the actual test for $\varepsilon \sim WN$ translates into a normality test, to verify that:

$$\sqrt{N}\hat{\rho}(\tau) \sim \mathcal{N}(0, 1) \quad \forall \tau > 0$$

The probability that $-\beta \leq \sqrt{N}\hat{\rho}(\tau) \leq \beta$ is given by the solution of the integral



The steps are the following:

1. Set a confidence interval $\alpha \in (0,1)$, possibly small. Compute β for which the area that underlies the two tails of $\mathcal{N}(0,1)$ is equal to α:

$$P(\sqrt{N}\hat{\rho} < -\beta, \sqrt{N}\hat{\rho} > \beta) = \alpha$$

2. Evaluate the number of samples of $\sqrt{N}\hat{\rho}(\tau) \notin [-\beta, \beta]$ and denote it with N_α.
If $\frac{N_\alpha}{N} < \alpha$, where N is the number of samples of $\hat{\rho}(\tau)$ available, then $\varepsilon(t)$ is assumed white, otherwise it is not.

6.10 Uncertainty in LS estimation

The Least Squares method is characterized by the following family of models:

$$M(\theta) : \hat{y}_\theta(t) = \theta'\varphi(t)$$

Suppose the system S is represented by:

$$S : y(t) = \theta^{0'}\varphi(t) + \eta(t)$$

Where $\theta^{0'}$ is the true parameter vector and $\eta \sim WN(0, \lambda^2)$. If S belongs to the family of models: $S = M(\theta^0)$ and there is a unique point of minimum, then we know that the estimate $\hat{\theta}_N \to \theta^0$ for $N \to \infty$.
We can also give an estimate of the uncertainty of the estimation $Var[\hat{\theta}_N - \theta^0] = \frac{1}{N}\lambda^2\bar{R}^{-1}$, which is a $n \times n$ matrix, where n is the total number of parameters to be identified. It tends to 0 as $\frac{1}{N}$ and the standard deviation tends to 0 as $\frac{1}{\sqrt{N}}$.
We introduce the gradient of the prediction error with respect to θ as

$$\psi_\theta(t) = \frac{\partial \varepsilon_\theta(t)}{\partial \theta}$$

Then we define:

$$\bar{R}(\theta) = E[\psi_\theta(t)\psi_\theta(t)']$$

which is an $n \times n$ matrix, too.
$\bar{R}(\theta)$ has to be evaluated at θ^0, which is feasible given the initial assumptions.
λ^2 is the variance of $\varepsilon_\theta(t)$ for $\theta = \theta^0$.
If we inspect the terms along the diagonal of the matrix $Var[\hat{\theta}_N - \theta^0]$, the i^{th} element is the variance of $\hat{\theta}_{N,i} - \theta_i^0$.

6.10.1 Evaluation of λ^2 and $\bar{R}$

We don't have the real values of λ^2 and $\bar{R}$, hence, we need to replace them with their surrogates obtained from the data.
$\varphi(t)$ can be constructed from the data for any $t = 1, 2, ..., N$. Then

$$\hat{R}_N = \frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi(t)'$$

We replace $\bar{R}$ with $\hat{R}_N$.
For λ^2 we can compute $\hat{\theta}_N$, the corresponding prediction error $\varepsilon_{\hat{\theta}_N} = y(t) - \varphi(t)'\hat{\theta}_N$.
From that, we can replace λ^2 with its estimate

$$\hat{\lambda}^2 = \frac{1}{N}\sum_{t=1}^{N}\varepsilon_{\hat{\theta}_N}(t)^2$$

6.11 LS procedure

Let's put all the pieces together and summarize the typical procedure to perform an LS estimation.

1. We have the model $\hat{y}(t) = \varphi(t)'\theta = \theta'\varphi(t)$
2. Compute the matrix $\sum_{t=1}^{N} \varphi(t)\varphi(t)'$
3. Check if it's invertible, if so compute

$$\hat{\theta_N} = \left[\sum_{t=1}^{N} \varphi(t)\varphi(t)'\right]^{-1} \sum_{t=1}^{N} \varphi(t)y(t)$$

4. Compute the prediction error of the estimated model

$$\varepsilon_{\hat{\theta}_N}(t) = y(t) - \varphi(t)'\hat{\theta}_N$$

5. Verify is $\varepsilon_{\hat{\theta}_N}(\cdot)$ is white: its covariance function should have values close to 0 for $\tau = 1, 2, ...$
6. If $\varepsilon_{\hat{\theta}_N}(\cdot)$ is a white noise, compute

$$Var\left[\hat{\theta}_N - \theta^0\right] = \frac{1}{N}\hat{\lambda}^2 \left[\frac{1}{N}\sum_{t=1}^{N} \varphi(t)\varphi(t)'\right]^{-1}$$

Example

$$S : y(t) = 1.2y(t-1) - 0.32y(t-2) + u(t-1) + 0.5u(t-2) + \eta(t)$$

with $\eta \sim WN(0,1)$ and $u \sim WN(0,4)$ (uncorrelated).
We generate 2000 data points and perform an estimation with an ARX(1,1), ARX(2,2), ARX(3,3).

ARX(1,1)

The values and the uncertainty of the two parameters are $\hat{a} = 0.932(0.6\%)$, $\hat{b} = 0.975(2.3\%)$.
The performance index is $J = 3.86$ but if we perform the whiteness test of $\varepsilon_{\hat{\theta}_N}(\cdot)$, it is not satisfied.

ARX(2,2)

Now we have four parameters $\hat{a}_1 = 1.2(1\%), \hat{a}_2 = -0.32(3\%), \hat{b}_1 = 0.98(1\%), \hat{b}_2 = 0.48(3\%)$.
$J = 0.99$ and the whiteness test is satisfied

ARX(3,3)

The values of the six parameters are $\hat{a}_1 = 1.19(2\%), \hat{a}_2 = -0.2(10\%), \hat{a}_3 = -0.019(68\%), \hat{b}_1 = 0.98(1\%), \hat{b}_2 = -0.49(5\%), \hat{b}_3 = -0.016(120\%)$.
$J = 0.97$ and the whiteness test is satisfied.
The parameters $\hat{a}_3$ and $\hat{b}_3$ very close to zero and with high uncertainty. This is a strong signal that the model may be too complex with respect to the data generation mechanism. Furthermore, there is no advantage in using a more complex model since the value of J is only slightly different.

Chapter 7

Model complexity selection

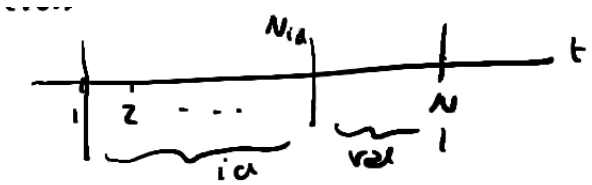
In general an $ARX(n, n)$ is a better fit to the data than an $ARX(n-1, n-1)$ model, but wee need to choose the best n in general.

7.1 Naive approach

If we simply compute the performance index for multiple increasing values of n we will see that is monotonically decreasing with n, but the decrease after a certain point may be very small and also it may not reflect in better performance for new unseen data.

7.2 Cross-validation

A better technique to find the best model complexity is by performing cross-validation, that is splitting the data points into two sets, one that is used to perform identification, while the other is used exclusively for the performance evaluation (validation set).



By using this approach, we are "wasting" some of the data, because they cannot be used in the identification process. Other alternative approaches to overcome this problem will be presented in the following sections.

7.3 Final Prediction Error (FPE)

The aim of the criterion is to evaluate

$$\bar{J}(\theta) = E[(y(t) - \hat{y}_\theta(t))^2]$$

which is the prediction error of model $M(\theta)$ for all the possible sequences of data.
We estimate the vector of parameters $\hat{\theta}_N$ through the usual minimization process. This value depends on the specific sequence of data used during the identification.
We want to compute the final prediction error:

$$FPE = E[\bar{J}(\hat{\theta}_N)]$$

which is a fair evaluation of the fitting capacity of model $M(\hat{\theta}_N)$ in the identification procedure. The minimum of FPE is appropriate to find the optimal complexity.

7.3.1 Derivation of FPE

Assume the following data generation mechanism:

$$S : y(t) = \varphi(t)'\theta^0 + \eta(t), \qquad \eta \sim WN(0, \lambda^2)$$

The model is

$$\hat{M}(\theta) : \hat{y}(t) = \varphi(t)'\theta$$

Then the prediction error is

$$\epsilon(t) = y(t) - \hat{y}(t) = \varphi(t)'\theta^0 + \eta(t) - \varphi(t)'\theta = \varphi(t)'(\theta^0 - \theta) + \eta(t)$$

Now we can compute:

$$\bar{J}(\theta) = E[\varepsilon(t)^2] = E[(\varphi(t)'(\theta^0 - \theta))^2] + E[\eta(t)^2] + 2E[\eta(t)\varphi(t)'(\theta^0 - \theta)]$$

Since the elements of the observations vector $\varphi(t)$ are $y(t-1), y(t-2), ...$ the expected value of the product between them and $\eta(t)$ is zero. Hence

$$\begin{aligned}\bar{J}(\theta) &= E[(\theta^0 - \theta)'\varphi(t)\varphi(t)'(\theta^0 - \theta)] + \lambda^2 \\ &= (\theta^0 - \theta)'E[\varphi(t)\varphi(t)'](\theta^0 - \theta) + \lambda^2 \\ &= (\theta^0 - \theta)\bar{R}(\theta^0 - \theta) + \lambda^2\end{aligned}$$

For the obtained expression of $\bar{J}(\theta)$ we can derive the expression of FPE, through the use of $Var[\theta^0 - \hat{\theta}_N] \simeq \frac{1}{N}\lambda^2\bar{R}^{-1} \Rightarrow \bar{R} = \frac{1}{N}\lambda^2 Var[(\theta^0 - \hat{\theta}_N)]^{-1}$.

$$\begin{aligned}FPE &= E[\bar{J}(\hat{\theta}_N)] \\ &= E[(\theta^0 - \hat{\theta}_N)'\bar{R}(\theta^0 - \hat{\theta}_N)] + \lambda^2 \\ &= \frac{\lambda^2}{N}E[(\theta^0 - \hat{\theta}_N)'Var[\theta^0 - \hat{\theta}_N]^{-1}(\theta^0 - \hat{\theta}_N)] + \lambda^2\end{aligned}$$

We define $\nu = \theta^0 - \hat{\theta}_N$ and then it can be demonstrated that $E[\nu' Var[\nu]^{-1}\nu]$ is equivalent to the number of parameters n. The FPE can be written as:

$$FPE = \frac{n}{N}\lambda^2 + \lambda^2$$

We replace λ^2 with its sampled version. The usual definition that multiplies the sum by $\frac{1}{N}$ is not the only one. Another possible definition is the following:

$$\hat{\lambda}^2 = \frac{1}{N-n}\sum_{t=1}^{N}\varepsilon(t)^2$$

and we finally obtain the expression of FPE

$$FPE = \frac{N+n}{N}\hat{\lambda}^2 = \frac{N+n}{N-n}\frac{1}{N}\sum_{t=1}^{N}\varepsilon(t)^2 = \frac{N+n}{N-n}J(\hat{\theta}_N)$$

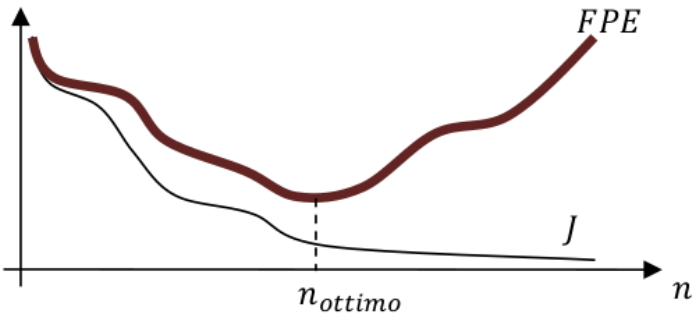
Through this approach we are giving a penalty to the models with high complexity. The FPE function is not monotonically decreasing, and the complexity corresponding to its minimum value can be chosen as complexity of the final model.

7.4 Akaike Information Criterion (AIC)

The concepts of the AIC are the same as in the FPE case. The only difference is the index used for the evaluation:

$$\begin{aligned}AIC = \ln(FPE) &= \ln\left(\frac{1+\frac{n}{N}}{1-\frac{n}{N}}J\right) = \ln(1+\frac{n}{N}) - \ln(1-\frac{n}{N}) + \ln(J) \\ &\approx \frac{n}{N} - (-\frac{n}{N}) + \ln(J) = 2\frac{n}{N} + \ln(J)\end{aligned}$$

The first term is the one regarding the complexity of the model, while the second is the one regarding the fitting of the data.



7.5 Minimal Description Length (MDL)

Another option is a more parsimonious one, because its minimum is achieved for lower orders with respect to the AIC case, which is the MDL:

$$MDL = (\ln N)\frac{n}{N} + \ln J$$

As complexity increases, the information needed to describe $M(\theta)$ is larger.

Chapter 8

Durbin-Levinson Algorithm

Suppose we are choosing the right complexity for an auto-regressive model AR(k) of order k. If we want to compute the parameters vectors for each value of k between e.g. 1 and 100, then we would need to invert 100 matrices of size $k \times k$, which is an expensive procedure.
The Durbin-Levinson algorithm is a recursive algorithm that allows to compute the solution of an AR($k+1$) starting from the solution of the AR(k).

8.1 From AR(1) to AR(2)

As an example, we will consider the passage from AR(1) to AR(2).
We have the following expression for the AR(2) model and we would like to obtain the values of $\tilde{a}_1$ and $\tilde{a}_2$, represented with a tilde to distinguish them from the AR(1) parameter a_1

$$\hat{y}(t) = \tilde{a}_1 y(t-1) + \tilde{a}_2 y(t-2)$$

we compute the covariance function $\gamma(\tau)$ of the $AR(1)$ model using the Yule-Walker equations.

$$\begin{aligned} \gamma(0) &= a_1\gamma(1) + \lambda^2 && (1)\\ \gamma(1) &= a_1\gamma(0) && (2)\\ \gamma(\tau) &= a_1\gamma(\tau-1), && \forall\tau > 1 \end{aligned}$$

Given $\gamma(\cdot)$, then

$$\begin{aligned} (2) \quad &\rightarrow \quad a_1 = \frac{\gamma(1)}{\gamma(0)}\\ (1) \quad &\rightarrow \quad \lambda^2 = \gamma(0) - a_1\gamma(1) \end{aligned}$$

Now we can pass to the AR(2) model:

$$\begin{aligned} \tilde{a}_1 &= a_1 - \tilde{a}_2\frac{\gamma(1)}{\gamma(0)}\\ \tilde{\lambda}^2 &= \lambda^2(1-\tilde{a}_2^2)\\ \tilde{a}_2 &= \frac{1}{\lambda^2}(\gamma(2) - a_1\gamma(1)) \end{aligned}$$

Note that $\tilde{\lambda} \leq \lambda^2$ and that $\tilde{a}_2 = 0$, i.e. AR(2) downgrades to AR(1), if $\gamma(2) = a_1\gamma(1)$, meaning that AR(1) fits $\gamma(2)$ as well. In that case we would have $\tilde{\lambda}^2 = \lambda^2$.
Another important observation is that from the second equation $\tilde{\lambda}^2$ can be negative, but it is not actually the case, because if we compute the transfer function of the AR(2) model

$$G(z) = \frac{1}{a - a_1 z^{-1} - a_2 z^{-2}} = \frac{z^2}{z^2 - a_1 z - a_2}$$

in the stationary case, the transfer function must be stable, i.e. the roots of $z^2 - a_1 z - a_2$ must be inside the unit disk, meaning that $|a_2| < 1$ because it is a product of the roots.
The Durbin-Levinson algorithm extends to any order of the AR.

8.2 From *k-1* to *k*

Consider the two models:

$$\mathrm{AR}(k-1) : y(t) = a_1^{(k-1)} y(t-1) + ... + a_{k-1}^{(k-1)} y(t-k+1) + \eta(t)$$

$$\mathrm{AR}(k) : y(t) = a_1^{(k)} y(t-1) + ... + a_k^{(k)} y(t-k) + \eta(t)$$

The parameter $a_k^{(k)}$ is called partial covariance coefficient:

$$PARCOV(\tau) = a_\tau^{(\tau)}$$

if the order of the "true" AR model is n then

$$PARCOV(\tau) = 0, \qquad \forall \tau > n$$

So, it can be used to derive the order of an appropriate auto-regressive model fitting the data.
It is parallel to the MA case, in which we have that

$$\gamma(\tau) = 0 \qquad \forall \tau > n$$

and we can determine the order by finding the value of τ for which the covariance function goes to zero. It can also used to determine if the model to be used should be an AR or an MA: if, at a certain point, the covariance function goes to zero before the partial covariance does, the process is a MA, while if the partial covariance goes to zero first, then it is an AR model should be used.

Chapter 9

Recursive Least Squares

All the algorithms seen so far are batch methods, that use all the data at once. Now we will see a recursive method that is able to update the estimate by adding new data. The latter methods help to overcome the limitation of when the data are coming some at a time.
The formula to compute the LS estimate in the batch version is

$$\hat{\theta}_t = \left(\sum_{i=1}^{t} \varphi(i)\varphi(i)' \right)^{-1} \sum_{i=1}^{t} \varphi(i)y(i)$$

We can define the first term sum as:

$$S(t) = \sum_{i=1}^{t} \varphi(i)\varphi(i)'$$

Now we want to find the relation between $\hat{\theta}_t$ and $\hat{\theta}_{t-1}$.
We can split the second factor of the first formula

$$\sum_{i=1}^{t} \varphi(i)y(i) = \underbrace{\sum_{i=1}^{t-1} \varphi(i)y(i)}_{S(t-1)\hat{\theta}_{t-1}} + \varphi(t)y(t)$$

We can replace it into the initial formula

$$\hat{\theta}_t = S(t)^{-1} \left[S(t-1)\hat{\theta}_{t-1} + \varphi(t)y(t) \right]$$

We can split $S(t)$ in the same way

$$S(t) = \sum_{i=1}^{t} \varphi(t)\varphi(t)' = \sum_{i=1}^{t-1} \varphi(i)\varphi(i)' + \varphi(t)\varphi(t)' = S(t-1) + \varphi(t)\varphi(t)'$$

We update the formula again to obtain the parameter updating equation:

$$\begin{aligned} \hat{\theta}_t &= S(t)^{-1} \left[(S(t) - \varphi(t)\varphi(t)')\hat{\theta}_{t-1} + \varphi(t)y(t) \right] \\ &= \hat{\theta}_{t-1} + S(t)^{-1}\varphi(t) \left[y(t) - \varphi(t)'\hat{\theta}_{t-1} \right] \\ &= \hat{\theta}_{t-1} + K(t)\varepsilon(t) \end{aligned}$$

where:

$$\begin{aligned} K(t) &= S(t)^{-1}\varphi(t) && \text{(gain)} \\ \varepsilon(t) &= y(t) - \varphi(t)'\hat{\theta}_{t-1} && \text{(prediction error)} \\ S(t) &= S(t-1) + \varphi(t)\varphi(t)' && \text{(auxiliary matrix updating)} \end{aligned}$$

Note that if $\varepsilon(t) = 0 \implies \hat{\theta}_t = \hat{\theta}_{t-1}$.
If $t \to \infty$:

- $S(t) \to \infty$ (monotonically increasing)
- $S(t)^{-1} \to 0$
- $K(t) \to 0$
- $\hat{\theta}_t \to \bar{\theta}$.

The last property is not always suitable, because the $\bar{\theta}$ may not be constant, but it may vary with time. A possible variant of the LS method to tackle this problem is through the introduction of a μ coefficient in the parameters updating that allows to give more importance to the new data with respect to the old one. In fact, μ is called forgetting factor and it is applied in the auxiliary matrix updating term:

$$S(t) = \mu S(t-1) + \varphi(t)\varphi(t)', \qquad \text{with } \mu \in (0, 1]$$

If $\mu = 1$ the method is the normal LS estimation.
If $\mu < 1$:

- at time $t : \mu^{t-t}\varepsilon(t) = \varepsilon(t)$
- at time $t-1 : \mu^{t-(t-1)}\varepsilon(t-1) = \mu\varepsilon(t-1)$
- at time $t-2 : \mu^{t-(t-2)}\varepsilon(t-2) = \mu^2\varepsilon(t-2)$

The performance index is updated, too, as:

$$J = \sum_{i=1}^{t} \mu^{t-i}\varepsilon(i)^2 = \sum_{i=1}^{t} \mu^{t-i}(y(t) - \varphi(t)'\theta)^2$$